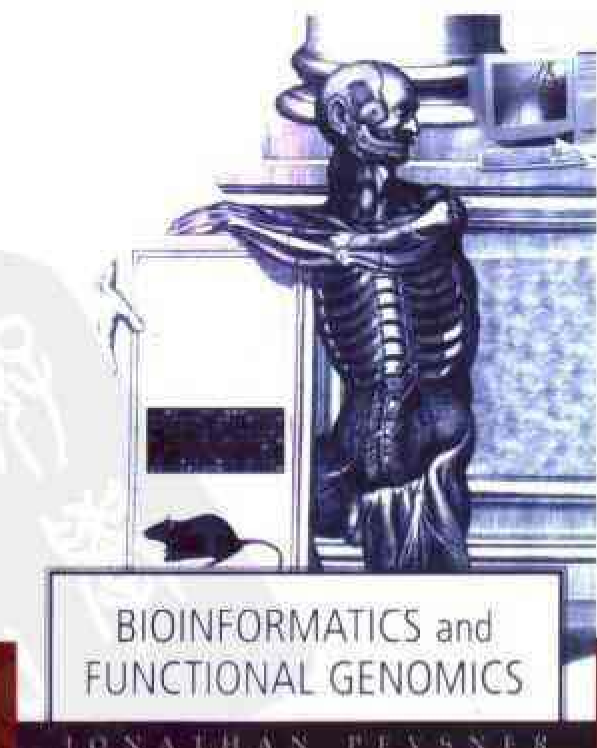
国外优秀科技著作出版专项基金资助

生物信息学与功能基因组学

BIOINFORMATICS and FUNCTIONAL GENOMICS

[美] 乔纳森·佩夫斯纳 著
(JONATHAN PEVSNER)

孙之荣 主译



Chemical Industry Press

化学工业出版社
现代生物技术与医药科技出版中心

VISIT…

LANZAROTE Caliente.com

生物信息学与功能基因组学

“对于生物学者，本书将给他们提供很有价值的指导，告诉他们如何利用生物信息学工具解决重要的生物学问题……对于计算/信息领域的学者……本书的介绍将使他们更加深刻地理解生物信息学的目的，并且帮助他们评价这一领域的存在的机会和最新的进展。”

—Russ B.Altman，斯坦福大学

翻开本书，您将会情不自禁地沉浸于其中……

本书内容全面，很好地将生物信息学与功能基因组学的理论和实践应用结合起来，非常重视生物信息学的具体应用技巧。另外，本书还具有如下特色：

- **作者权威**。本书作者是霍普金斯医学院教授，在国际上有很高的声誉和权威性。
- **实践性强**。每一章都有问题集、与生物信息学有关的web操作训练以及相应的web链接，还列出了可以免费获取的生物信息学软件和作者推荐的读物。
- **范例典型**。本书从头到尾都用一个基因和它的蛋白产物作为例子进行讲解。这个蛋白是视黄醇结合蛋白（retinol-binding protein，RBP）。
- **图文并茂**。本书在论述的同时配以大量的图片，直观、形象、通俗易懂。

即使您没有任何计算机背景，同样可轻松地阅读本书。

ISBN 7-5025-8383-1

9 787502 583835 >

销售分类建议：生物/生命科学

ISBN 7-5025-8383-1　定价：95.00元

国外优秀科技著作出版专项基金资助

生物信息学与功能基因组学

BIOINFORMATICS and FUNCTIONAL GENOMICS

[美] 乔纳森·佩夫斯纳 著
(JONATHAN PEVSNER)

孙之荣 主译

化学工业出版社
现代生物技术与医药科技出版中心
·北京·

图书在版编目（CIP）数据

生物信息学与功能基因组学/[美] 乔纳森·佩夫斯纳（Pevsner J.）著；孙之荣主译. —北京：化学工业出版社，2006. 5

书名原文：Bioinformatics and Functional Genomics

ISBN 7-5025-8383-1

Ⅰ. 生… Ⅱ. ①佩…②孙… Ⅲ. ①生物信息论②基因组-研究 Ⅳ. ①Q811. 4②Q343. 2

中国版本图书馆 CIP 数据核字（2006）第 020646 号

Bioinformatics and Functional Genomics/by Jonathan Pevsner

ISBN 0-471-21004-8

Copyright©2003 by John Wiley & Sons, Inc. All rights reserved.

Authorized translation from the English language edition published by John Wiley & Sons, Inc.

本书中文简体字版由 John Wiley & Sons，Inc. 授权化学工业出版社独家出版发行。

未经许可，不得以任何方式复制或抄袭本书的任何部分。

北京市版权局著作权合同登记号：01-2005-2537

生物信息学与功能基因组学

[美] 乔纳森·佩夫斯纳 著

孙之荣 主译

责任编辑：邵桂林 郎红旗

责任校对：陶燕华

封面设计：胡艳玮

*

化学工业出版社
现代生物技术与医药科技出版中心 出版发行

（北京市朝阳区惠新里 3 号 邮政编码 100029）

购书咨询：(010)64982530

(010)64918013

购书传真：(010)64982630

http://www.cip.com.cn

*

新华书店北京发行所经销

北京永鑫印刷有限责任公司印刷

三河市万龙印装有限公司装订

开本 787mm×1092mm 1/16 印张 45¾ 字数 1130 千字

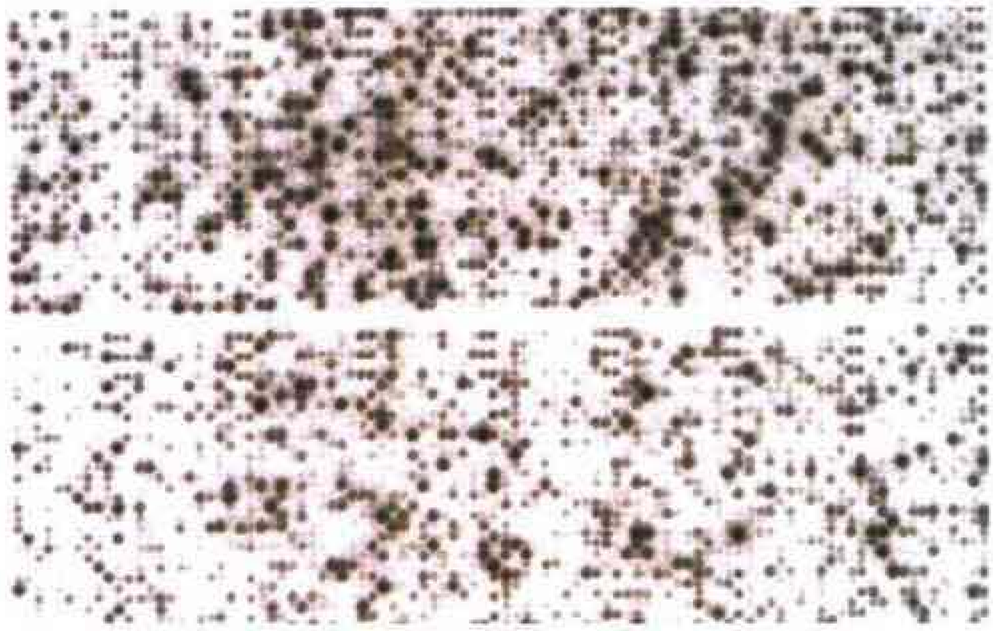
2006 年 6 月第 1 版 2006 年 6 月北京第 1 次印刷

ISBN 7-5025-8383-1

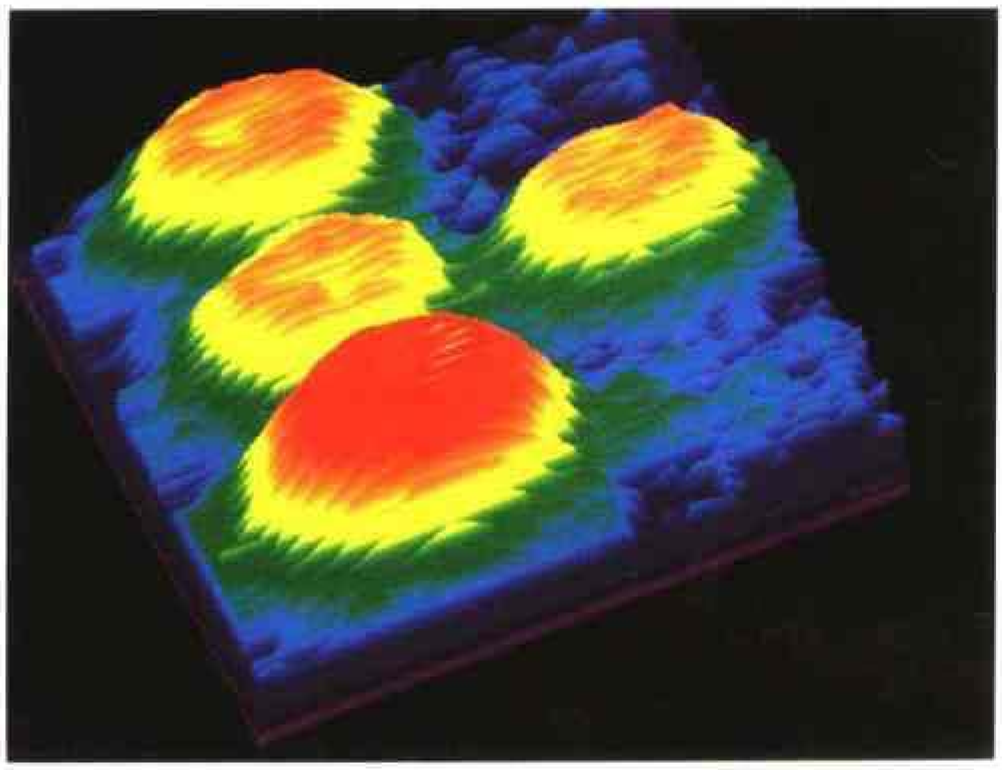
定 价：95.00 元

版权所有 违者必究

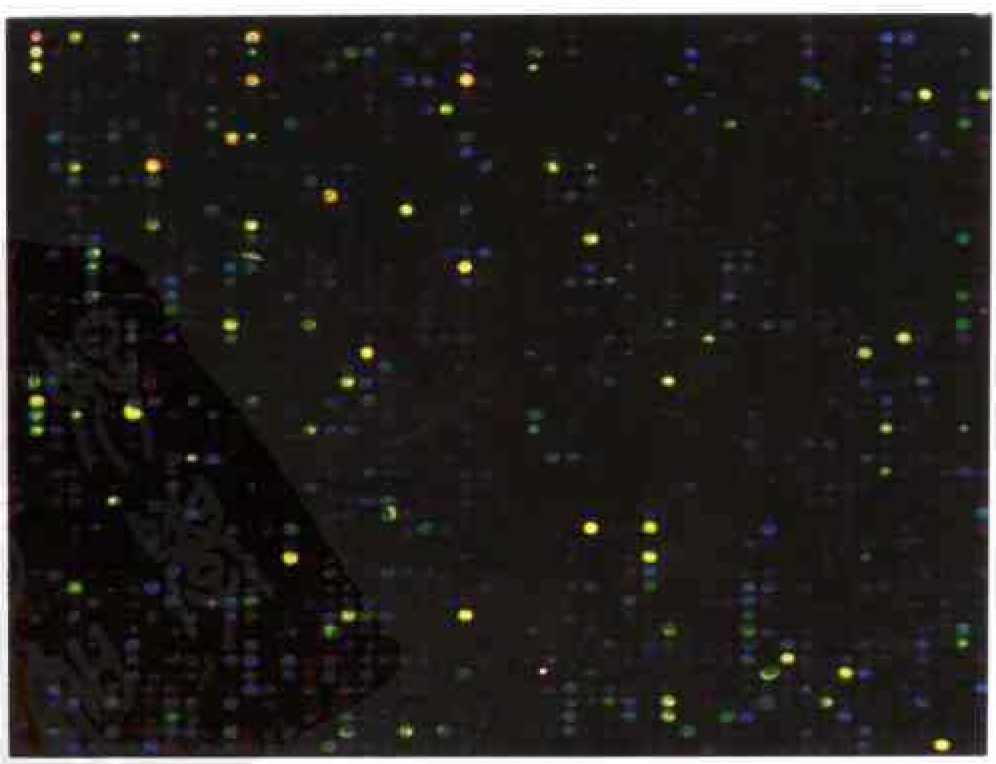
该书如有缺页、倒页、脱页者，本社发行部负责退换



(a)



(b)



(c)

图 6.20 Research Genetics公司的GeneFilter微阵列（探针为从唐氏综合征患者死后脑部海马区得到的cDNA）

（a）微阵列上共有5000种cDNA。基因在所有阵列上的分布必须随机化；

（b）用NIH图像软件看到的6个信号点。图像分析软件必须详细说明每个信号的性质，包括一个强信号（图中左下角）“溢出”到弱信号（右下角）的可能性；

（c）NEN Perkin–Elmer公司的一种微阵列（MICROMAX，有2400个基因），用和图6.19中同一个瑞特综合征患者和相应对照组的脑组织样本做探针。这里使用的是经荧光标记和竞争性杂交的样本

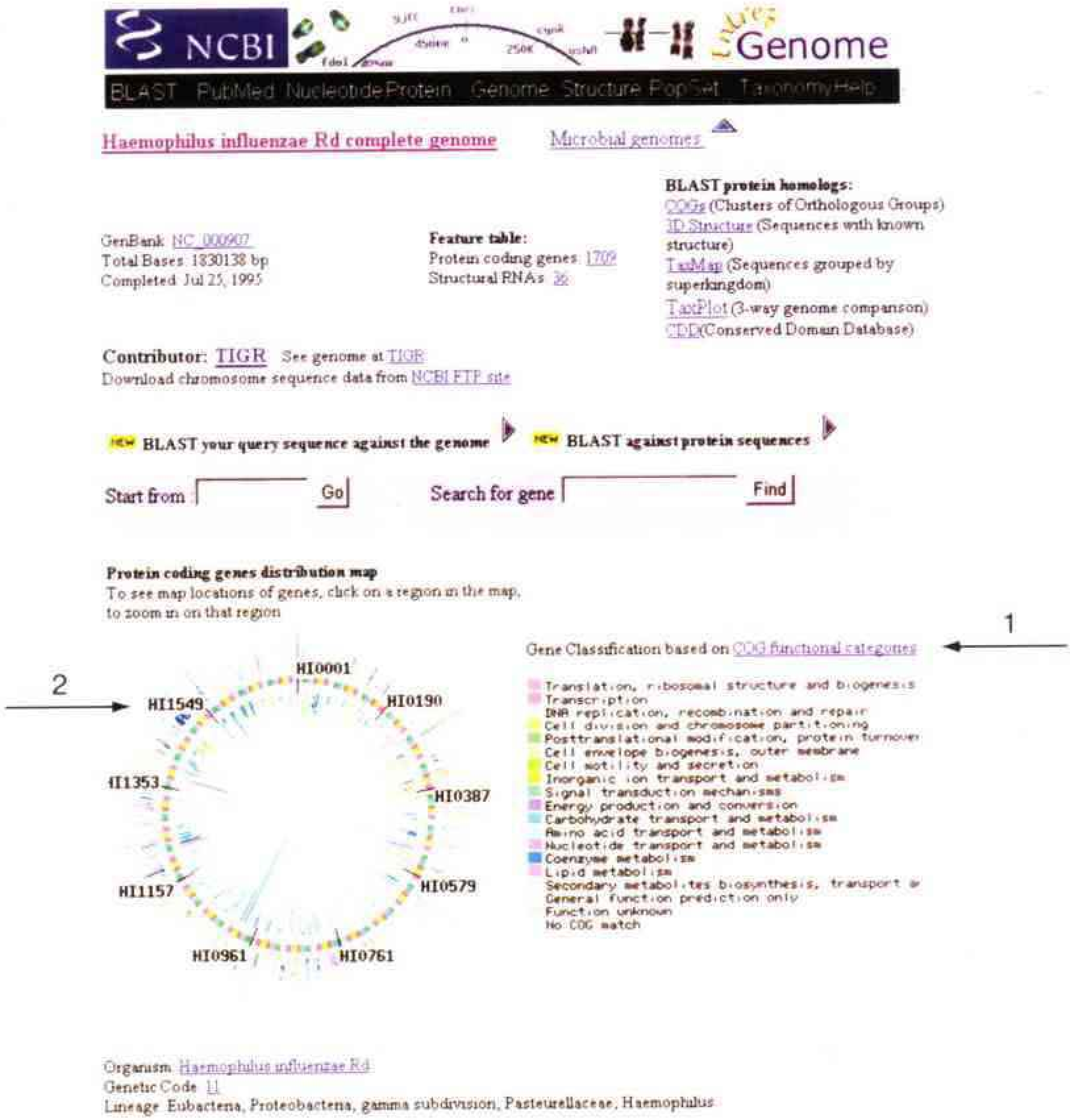


图 12.9 第一个完成基因组测序工作的独立生活的生物体 *H. influenzae* Rd的Entrez基因组记录。这个记录是从Entrez基因组资源中点击其左侧栏的“bacteria（细菌）”得到的。记录的顶端包含了基因组的编号以及大小等信息。整个核苷酸序列都可以从这里下载。右上方是关于这个基因组所编码的1709个蛋白质的一些资源。从这里可以进行 *H. influenzae* 的BLAST搜索。记录的主要部分是由一个代表这个基因组的彩色圆环和一个基于COG（见下文）的功能分类组成的。环形图表被点击以后（图12.10）将显示这个基因组的基因和蛋白质的详细信息。底部是这个细菌的基因代码以及分类学起源。这个记录还包含了参考文献（图中没有显示出来），包括基因组研究所Fleischmann等在1995年的最初的基因组序列报告

译者序

后基因组时代的生物学研究对每一个生物学工作者来说都充满了挑战。随着人类基因组测序工作的初步完成，越来越显现出生物信息学的重要作用。目前，虽然生物信息学已经在国内外得到了蓬勃的发展，但是国内相关书籍的水平还十分参差不齐，至今仍没有一本非常翔实的中文教材。与此同时，很多生物学研究者特别是生物信息学工作者都迫切需要一本能够详尽介绍生物信息学概要的中文教材类用书，而《生物信息学与功能基因组学》的出现恰可部分满足这种需求。此书内容翔实、语言通俗、图文并茂，对目前生物信息学的各个主要研究方向讲解详细而清楚，引用了国外生物信息学领域专家最新的研究成果。特别是第3篇对各个模式生物基因组序列所作的单独讲解很有特色，无论是对于生物信息研究者还是生物学研究者都将会是十分实用的。书后的生物信息学专业名词术语表设计，便于读者在阅读和平时学习及工作中作简单查阅。我们将此书整理翻译出版并介绍给广大的生物学研究工作者，希望能对他们的研究工作有所帮助，同时此书对从事生物信息学研究和学习的大学本科生和研究生也是一本很好的参考教材。

《生物信息学与功能基因组学》涵盖了现代生物学的许多新的研究领域，对功能基因组有比较详细的介绍，提供了一些重要的生物学资源。本书将生物信息学与功能基因组学结合在一起讲述，更清楚地描述了功能基因组时代生物信息学的新发展。全书分为3篇，第1篇着重分析了当前web数据库中的DNA、RNA和蛋白质序列资源；第2篇着重在基因组层次上对蛋白质和RNA进行了分析；第3篇着重对多个具体的已经测序的基因组完整序列进行了详尽的分析；书后的附录部分还附带了生物信息学专业名词术语表。

本书的翻译工作由清华大学生物科学与技术系生物信息学与系统生物学研究所的师生集体完成。参加本书翻译校对工作的有李立博士、郑家顺博士、夏学峰、沈金城、孟祥波、王世雷、周云、袁麟、李薇、黄泥、张松、陈虎、吴明、黄波、陈香、梁灵等。全书由孙之荣教授负责组织和审校。希望本书的翻译和出版将对后基因组时代我国生物学研究的发展及对生物信息学研究人才的培养能够提供良好的帮助和促进作用。

孙之荣

2006年1月于清华园

原序

近10年来生物信息学的成长基于一个简单的原因：数据。逐渐积累的大量实验数据驱使我们开发数据的表示方法和算法来管理这些信息。生物学正在发生一场革命，如果计算的革命不同时伴随着生物学的革命，那么计算机能为生物学服务的内容就比较有限，如同20世纪80年代中期那样：一些科学家（如结构生物学家）利用计算机来进行模型优化、结构模拟和可视化，而另一些热心的科学家会醉心于管理专门的“精品店”式的数据库，大多数生物学家则仅会偶尔地使用计算机进行文献搜索、结果的图形化、定性模型的描述，以及少量的序列搜索和比较。

但是，生物学发生了革命。实验测定生物分子（蛋白质和核酸）序列能力的增长超出许多人的预期。在20世纪90年代早期我们已经清楚我们需要改进方法来储存、分析和传播这些生物学信息。科学团体启动人类基因组计划（和相关的测序计划）的决定使得那些懂得如何分析数据和拥有发明新方法的能力的科学家们大有用武之地。其结果是，来自于数学、统计学和计算机科学的研究者们导致了第一次突破，这些发现可以用于帮助拼接基因组序列、创建序列数据库以及开发序列分析的快速搜索和比较算法。

在早期，从事生物信息学的人员来自于各个领域。研究某些特定问题的生物学家进入了计算领域：X射线结晶学、NMR波谱学、系统发生学和群体（或统计）遗传学。同时，三维结构计算也吸引了一些计算机科学家的注意，他们的兴趣在于计算机图形学和计算几何。对字符串分析算法感兴趣的计算机科学家发现序列分析也是富于挑战和吸引人的课题。近年来，一些统计学家钟情于分析富有噪声但又有意义的微阵列数据。这些先驱研究者们创造了大量的方法。一些方法用途广泛，为常见的问题提供了解决方法（如BLAST），一些能成功用于解决特定的生物学问题，但应用于更广泛领域的能力则有限，一些方法虽然被阐述得很有意义，但不能用于解决生物学问题。

基因组测序计划从两个方面扩展了生物信息学的覆盖范围。首先，它们提供了研究的关键材料，使得该领域有了自己的使命并且能够成长起来。另一方面，生物学家们意识到基因组计划的成果能通过计算分析的方法为他们所用。因此，在这个领域发展的早期就得到了大量的来自生物学团体的关键支持。此种情形促使了一些支持生物信息学研究和培训计划的产生，并且在生物学家和生物信息学家之间建立了良好的关系，生物信息学家们及时地开发了一系列有用的工具，因此建立了学科互动和产出的可持续模式。

生物信息学的未来在有些方面是清楚的，在另一些方面则是模糊的。毫无疑问，许多生物学家赞成生物学研究中的“高通量”方法，并且正在发明各种新方法来获取数量巨大的有用信息。在发生快速测序方法的革命之后，其他大规模测量数据的方法也随之而来。迄今影响最大的是测量细胞内mRNA表达量的微阵列技术。实验室的小型化也加速了我们其他许多实验手段和能

力的发展。例如，使用二维凝胶电泳和质谱技术分析全细胞的蛋白质（和它们与翻译后修饰的蛋白质的区别）提供了海量数据。生物学家产生数据的能力和兴趣是没有限制的，因此，从这个方面来说，生物信息学是不用担心的。事实上，数据积累的速度将会加速，所以模型会变得更复杂，也会需要模拟和可视化方面更进一步的创新。

另一方面，生物信息学将以何种方式成熟起来，也是存在疑问的。一些人认为"工具开发者"和"工具使用者"之间是存在重要区别的，开发者接受技术学科的培训，并针对由使用者描述的一般需求来产生可靠的方法。他们甚至位于专门的"生物信息学"系，并和其他的工具开发者们相互联系与合作。使用者们能理解这些工具，并合作参与开发，但是他们的主要兴趣在于帮助他们解决问题。另外一些人则认为生物信息学将以一种更加整合的方式并入生物科学的框架。下一代生物学家将会拥有定量和计算领域的知识和技巧，它们将会是在生物学研究中碰到需要解决的问题而发展起来的新的表示方法和算法的主要来源。开发者和使用者之间的区别将消失，因为生物学家研究解决问题的方法将会集中于生物信息学方法和其他方法，例如实验设计。通常，答案位于中间的某处。开发工具所需要的专门的定量和计算方法来自于生物学之外，但是生物学家们会在他们的工作中使用这些技术学科的知识和方法。这本书就是为这种整合而迈出的重要一步。在《生物信息学和功能基因组学》中，Jonathan Pevsner 写下了令人感兴趣的文字。

表面上，这本书是针对那些想解决问题的生物学家的，例如，它的组织是基于基本的生物学法则，即从 DNA 到 RNA 再到蛋白质，同时也从单基因分析到多基因分析再到基因组分析；对相关方法的描述是为了帮助用户理解这些算法的关键特性，而不涉及不必要的实现上的技术细节。但是，这本书不止如此。它不仅仅描述算法，以作为解决问题时的手册之用，它也提供了近 10 年来生物学上的许多重要发现和突破。有一点不得不说：生物信息学方法是基因组革命的一部分，因此在解释生物信息学工具时，必须描述开发这个工具的生物学背景和问题。因此，这本书不仅对生物学家（他们希望解决问题）而且对计算科学家（他们希望了解生物信息学处理的生物学问题，哪些已被解决，哪些重要问题还未被解决）来说都是有用的。

要合适地开发一个工具并提供有意义的解释需要一定的生物学知识深度，这一点使我产生深刻印象。例如，对分子系统发生和进化的讨论（第 11 章）提供了分子进化整个领域的优秀教程，它和对计算工具的描述同等有用。类似地，对基因表达的讨论（第 6 章、第 7 章）非常好地概述了现在的基因表达模型，例如这些数据如何用于提出假设，以及什么情况下这些假设是不足的。最后，对基因组的综述（第 12～第 17 章）极好地概括了我们面对的各个物种的生物学挑战，阐明了模式生物的重要性，也因此清晰地说明了我们需要处理的重要信息学问题。

因此，留给我们的是激动人心的新发现，对生物学界和计算/信息科学界都是如此。对生物学家而言，我们可以为针对重要的生物学问题开发重要工具提供宝贵的指导意见。一段写作清晰的文字需要附上观点、不足、疑问、问题、测验和参考资料（在线的和离线的），其目的是为了让一个严肃的科学工作者能快速地获取在生物信息学领域中分析当前问题的领先的方法。对计算/信息科学界而言，我们需要对这些方法拟解决的关键问题做一概要的描述，如为什么这一问题比较困难，现有的方法解决到怎样的水平。因此，提供关于生物信息学的生物学动因的重要介绍，也可用来评估生物信息学的发展机会和现有进度。基于这些理由，我热心地推荐此书，同时希望您利用本书既能理解当前生物信息学工具的状态（和如何熟练地使用它们），也能了解它们背后的生物学问题（包括解决的和留下的）。

Russ Altman

斯坦福大学

前言

1. 本书的起源

这本书源于几年以前我在 Johns Hopkins School of Medicine 为介绍生物信息学和基因组学课程准备的讲义。第一次课包括大约 70 位研究生和数百名听众，他们中有博士后研究员、技术员、本科生和教职人员。参加课程的人员来自于各个领域——遗传学、神经科学、免疫学和细胞生物学的学生，对特殊疾病感兴趣的临床医生，统计学家和计算机科学家，病毒学家以及微生物学家。他们有一个共同的兴趣，就是想知道如何使用计算机来解决生物学问题。对生物信息学，我简单地定义为计算机科学和分子生物学的交叉。这个新兴的领域依赖于使用计算机算法和计算机数据库来研究蛋白质、基因和基因组。功能基因组学就是使用基因组范围的实验和计算方法来研究基因的功能。

2. 比较

本质上说，生物信息学的研究领域是关于比较研究的领域。本书的前 1/3 我们将学习如何从数据库中取得 DNA 或蛋白质序列，然后成对地相互比较它们或者搜索整个数据库。对一名对一条特定基因感兴趣的学生来说，一个自然的问题是"其他什么基因（或蛋白质）和我这个基因相关?"。

在本书的前 1/3 部分，我们沿着从 DNA 到 RNA（基因的表达）再到蛋白质的路线进行描述，并进行一系列的比较。我们会比较两个细胞株的基因表达，它们是经过药物处理和未经处理过的两个细胞株，或者是野生型小鼠心细胞和基因敲除后的小鼠心细胞，或者处于发育不同阶段的蛙细胞。这些比较会扩展到蛋白质的世界，那里我们将应用蛋白质组学工具处理多种生理条件下的复杂生物学样本。多个相关的 DNA 和蛋白质序列比对是另一种形式的比较。这种相关性可以用系统发生树来进行可视化。

本书的后 1/3 部分论述了生命之树（the tree of life），这提供了另一个层次上的比较。哪种人类免疫缺陷病毒（HIV）对我们有威胁，我们如何能通过比较各种 HIV 亚型来帮助我们开发疫苗？蚊子和果蝇是怎么相关的？脊椎动物（如鱼和人）共享哪些基因，哪些基因是各系统发生谱系特有的？

我相信是这些各种不同的比较使得新兴的生物信息学和基因组学区别于传统的生物学领域。生物学总是关注于比较。书中，我提到了 19 世纪的生物学家，如 Richard Owen、Ernst Haeckel 和 Charles Darwin，他们在生物体的层次上进行了比较研究。我们准备解决的问题实质上没有改变。我们仍在寻求对生物学的统一概念的更完整的理解，例如生命的组成部分（如蛋白质和基

因）的组织、复杂生物系统的行为和生命进化的连续性。所改变的是我们寻求这种完整理解的方法。这本书描述了富有各种基因的原始信息和基因产物的数据库，以及用来分析这些数据的工具。

3. 人类疾病的挑战

我将学生当作一名分子生物学者和神经科学者进行培训。我的实验室研究儿童脑部疾病（如 Down 综合征、孤独症和铅中毒）的分子基础，它位于 Kennedy Krieger Institute，是一家儿童发育疾病的医院（你可以从 http://www.kennedykrieger.org 获取更多的信息）。每年有超过 1 万名病人访问该机构。医院开设了各种儿童疾病（包括语言障碍、饮食失调、孤独症、智力迟钝、脊柱分裂以及脑外伤）的临床部门。其中一些是常见的疾病，如 Down 综合征（大约 700 名新生儿中有 1 名受影响）和智力迟钝；也有一些罕见的疾病，如 Rett 综合征或者肾上腺脑白质营养不良（*adrenoleukodystrophy*）。

现在，世界公共数据库中储存的碱基数量已达到近百亿，这在第 2 章中会有所描述。我们也已经得到了人类基因组的序列，而且，自 1995 年以来有数百个基因组已被测序。贯穿本书，你将会沿着科学进步的路线学会如何对 DNA 测序，研究它的 RNA 和蛋白质产物。当今，科学进步的脚步让人眼花缭乱。

但是，与此同时我们对人类疾病仍然所知甚少。有上千种疾病是由于单个基因的一个缺陷导致的病态现象。不过即使我们发现了疾病中所发生缺陷的基因，例如囊肿纤维症（cystic fibrosis）、肌营养不良、肾上腺脑白质营养不良和 Rett 综合征，但要做到有效治疗或治愈还是很困难。与单基因疾病相比，复杂疾病更常见，如孤独症、抑郁症和智力迟钝，它们可能是由多基因突变引起的。与遗传疾病相比，感染性疾病又更常见。我们对于为什么一个病毒株仅感染人类而与之紧密相关的另一个病毒株仅感染黑猩猩所知更少。我们不明白为什么一种菌株是致病的而另一菌株又是无害的。我们还不了解如何开发有效的疫苗来对付真核致病原，包括从原生动物（例如恶性疟原虫*Plasmodium falciparum* 导致疟疾）到寄生性的线虫。

近来，生物信息学工具的开发导致在这些领域取得进步的前景是鼓舞人心的。我们仅仅刚刚开始理解致病体和易感性宿主的遗传基础。我们的希望是，在生物信息学数据库中迅速积累起来的信息，能通过研究转化成对人类疾病和一般生物学机制的洞察理解。

4. 对读者的提醒

本书描述了和生物信息学与功能基因组相关的 1000 个以上的网址。所有这些地址会随时间

变化（并且一些会消失）。为了让这些网址链接与时俱进，一个相伴的网址（http://www.bioinfbook.org/）基本上维护了书中所有的网址链接，并按书中的章节排列。我们尽我们所能随时间维护这些网址。每个月，我们用一个程序扫描所有链接，并且如有必要就更新它们。

另一个网址是针对教师的，包含了对问题的详细解答（参见 http://www.wiley.com/）。

5. 致谢

这本书的写作过程是一个很好的学习经历。非常感谢许多对此做出贡献的人们，尤其是 Kennedy Krieger 研究所和 Johns Hopkins 医学院，这里的知识氛围异常浓厚。书中的各章节是由一个介绍生物信息学课程的讲义发展而来的。学习该课程的前三年中，Johns Hopkins 医学院的教员为 Jef Boeke（负责酵母功能基因组学）、Aravinda Chakravarti（负责人类疾病）、Neil Clarke（负责蛋白质结构）、Kyle Cunningham（负责酵母）、Garry Cutting（负责人类疾病）、Rachel Green（负责 RNA）、Stuart Ray（负责分子系统发生学），以及 Roger Reeves（负责人类基因组）。他们对这些领域的见解令我获益匪浅。

我由衷地感谢本书的许多审阅者，包括一群匿名的审阅者，他们提供了相当有建设性和详细的建议。审阅本书的包括 Russ Altman、Christopher Aston、David P. Leader 和 Harold Lehmann（各个章节），Conover Talbot（第 2 章），Edie Sears（第 3 章），Tom Downey（第 7 章），Jef Boeke（第 8 章和其他一些章节），Michelle Nihei 和 Daniel Yuan（第 8 章），Mario Amzel 和 Ingo Ruczinski（第 9 章），Stuart Ray（第 11 章），Marie Hardwick（第 13 章），Yukari Manabe（第 14 章），Kyle Cunningham 和 Forrest Spencer（第 15 章）和 Roger Reeves（第 16 章）。Kirby D. Smith 阅读了第 18 章并且对其他各章也提出了见解。每位同事都花了大量的时间和精力来帮助提高文中的内容，每一位都堪称导师。许多学生也阅读了书中一些章节，我要提及的有 Rong Mao、Ok-Hee Jeon 和 Vinoy Prasad。我特别感谢 Mayra Garcia 和 Larry Frelin，他们在我写作的过程中提供了宝贵的帮助。我还要感谢 John Wiley & Sons 的编辑 Luna Han 对我的鼓励。

我也要感谢 Gary W. Goldstein（Kennedy Krieger Institute 的院长）和 Solomon H. Snyder（Johns Hopkins 神经科学系主任），他们都提供了鼓励和帮助，并且让我在从事研究的同时有机会来写这本书。

我感谢我的家人，因为他们提供了关爱和支持；同时也感谢 N. Varg、Kimberly Reed 和 Charles Cohen。最后，我要感谢我的未婚妻 Barbara Reed，是因为她的耐心、忠诚和爱。

主题的背景介绍。对一些高水平的学生，每章最后引用了一些关键的研究文献。这些文章是技术性的，阅读这些文章能加深对所在章节内容的理解。推荐读物部分也包括了综述文章。

1.6 重要的生物信息学网址

生物信息学领域的研究严重依赖于因特网，通过它来访问序列数据、获得分析分子数据的软件，并且整合生物学相关的不同信息和资源。我们将提供一系列的网址。开始，我们集中于三个主要的存放 DNA 和蛋白质数据的公共数据库（表 1-1）。在第 2 章，我们以 National Center of Biotechnology Information（NCBI）开始（GenBank 位于 NCBI 上）。NCBI 网站提供了许多其他生物信息学相关工具。然后我们将慢慢地介绍欧洲生物信息学研究所（European Bioinformatics Institute，EBI）网站，该网站中有一补充的 DNA 数据库（the European Molecular Biology Laboratory database，EMBL）。我们也会介绍日本的 DNA 数据库（DNA Database of Japan，DDBJ）。GenBank、EMBL 和 DDBJ 的研究团队每天更新共享序列数据。生物信息学学科的一个普遍特征是许多数据库是相互紧密连接的。

表 1-1 三个主要的生物信息学 Web 服务器作为 DNA 和蛋白质序列数据的中心仓库（这些将在第 2 章介绍）

资 源	描 述	URL①
日本 DNA 数据库(DDBJ)	和信息生物学中心(center for information biology)相关	http://www.ddbj.nig.ac.jp/
欧洲生物信息学研究所(EBI)	维护 EMBL 数据库	http://www.ebi.ac.uk/
美国国立生物技术信息中心(NCBI)	维护 GenBank	http://www.ncbi.nlm.nih.gov/

① 统一资源定位器（uniform resource locator）。

我们将在本书所有章节介绍共数百个生物信息学相关的网址。表 1-2 列举了一些服务器，提供数据库服务，也提供许多序列分析相关的程序。表 1-3 列出了一些地址，它们提供生物信息学资源的链接。我们现在就为那些希望探究当前可用的生物信息学资源类型的人们列出这些地址。

表 1-2 其他生物信息学 Web 服务器

资 源	描 述	URL
Centre for Molecular and Biomolecular Informatics	来自 Nijmegen 大学	http://www.cmbi.kun.nl/
ExPASy(Expert Protein Analysis System)	Swiss Institute of Bioinformatics 的蛋白质组服务器	http://www.expasy.org/
GENESTREAM	Institut de Genetique Humaine, Montpellier	http://www2.igh.cnrs.fr/
GenomeNet	位于日本京都	http://www.genome.ad.jp/
INFOBIOGEN	位于 Montpellier	http://www.infobiogen.fr/page_accueil_en.html
Oak Ridge National Laboratory (ORNL)	位于 Tennessee	http://compbio.ornl.gov/
Protein Information Resource (PIR)	美国国立生物医学研究基金会的一个部门	http://pir.georgetown.edu/
The Wellcome Trust Sanger Institute	位于英国 Cambridge 的基因组研究中心	http://www.sanger.ac.uk/
The Institute for Genomic Research(TIGR)	位于美国 Rockville, Maryland	http://www.tigr.org/

注：每个网站都包含了可访问的许多软件工具、研究计划、引用文献，以及其他生物信息学相关信息。

表 1-3 提供有用链接的生物信息学站点（提供生物信息学资源链接的网站）

资源	描述	URL
Amos' WWW links page	来自 ExPASy	http://www.expasy.ch/alinks.html
DBCAT, The Public Catalog of Databases	来自 INFOBIOGEN	http://www.infobiogen.fr/services/dbcat/
European Molecular Biology Network	各个欧洲站点	http://www.embnet.org/
Human Genome Most Used Links	来自美国 Los Alamos 国家实验室	http://www-ls.lanl.gov/HGhotlist.html

1.7 推荐读物

生物信息学领域的概览可参见由 Mark Gerstein 和他的同事（Luscombe 等，2001）以及 Claverie 等（2001）撰写的文章或著作。Kaminski（2000）也介绍了生物信息学，并附有访问网站的建议。Russ Atlman（1998）讨论了生物信息学和医药的相关性，而 David Searls（2000）则介绍了研究基因组的生物信息学工具。

参 考 文 献

Akerstrom, B., Flower, D. R., and Salier, J. P. Lipocalins: Unity in diversity. *Biochim. Biophys. Acta* **1482**, 1-8 (2000).

Altman, R. B. Bioinformatics in support of molecular medicine. *Proc. AMIA Symp.*, 53-61 (1998).

Bishop, R. E. The bacterial lipocalins. *Biochim. Biophys. Acta* **1482**, 73-83 (2000).

Boguski, M. S. Bioinformatics. *Curr. Opin. Genet. Dev.* **4**, 383-388 (1994).

Claverie, J. M., Abergel, C., Audic, S., and Ogata, H. Recent advances in computational genomics. *Pharmacogenomics* **2**, 361-372 (2001).

Flower, D. R., North, A. C., and Sansom, C. E. The lipocalin protein family: Structural and sequence overview. *Biochim. Biophys. Acta* **1482**, 9-24 (2000).

Frankel, A. D., and Young, J. A. HIV-1: Fifteen proteins and an RNA. *Annu. Rev. Biochem.* **67**, 1-25 (1998).

Goodman, N. Biological data becomes computer literate: New advances in bioinformatics. *Curr. Opin. Biotechnol.* **13**, 68-71 (2002).

Henikoff, S. Beyond the central dogma. *Bioinformatics* **18**, 223-225 (2002).

Ideker, T., Galitski, T., and Hood, L. A new approach to decoding life: Systems biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343-372 (2001).

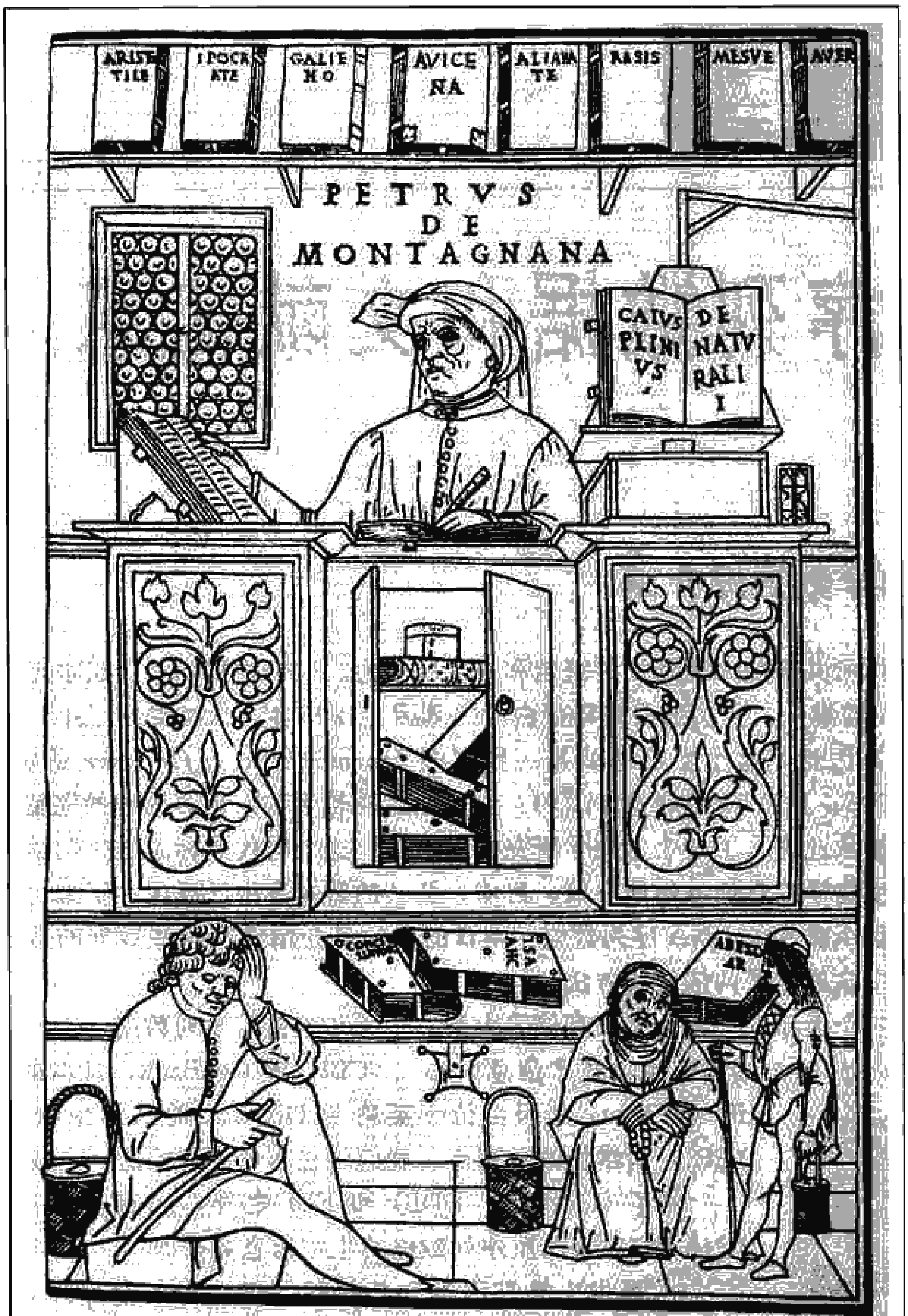
Kaminski, N. Bioinformatics. A user's perspective. *Am J. Respir. Cell Mol. Biol.* **23**, 705-711 (2000).

Luscombe, N. M., Greenbaum, D., and Gerstein, M. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf. Med.* **40**, 346-358 (2001).

Newcomer, M. E., and Ong, D. E. Plasma retinol binding protein: Structure and function of the prototypic lipocalin. *Biochim. Biophys. Acta* **1482**, 57-64 (2000).

Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734-740 (1997).

Searls, D. B. Bioinformatics tools for whole genomes. *Annu. Rev. Genomics Hum. Genet.* **1**, 251-279 (2000).



第 2 章介绍了如何获取生物分子的数据。这张图选自 Johannes de Ketham 著的《Fasciculus medicinae》，这本书的第一版于 1491 年出版．这张画展现了三个病人坐在 Petrus de Montagnana 办公室前的情景．Petrus de Montagnana 是一名医生，也是 Padua 医学院的一名教授 (Singer 于 1925 年根据 1493 年的图画重新绘制了这幅画)。在画中，我们看到 Petrus de Montagnana 拥有许多书籍，其中 Pliny 著的《Nature History》摊开着放在了一张桌子上。上面的书架上陈列着亚里士多德、Hippocrates、Galen、Ibn Sina、Haly Abbas、Rhazes、Mesue 和 Averroes 的著作，底板上则放着 Pietro d'Abano（《Conciliator》)、Issac 和 Avenzoar 的作品。Johannes Gutenberg 于 1450 年开创了他的印刷出版业，到 1500 年，仅在威尼斯城就已经出版了 200 万部图书（Sarton，1955）。从某种意义上讲，我们处在目前的基因组时代（从 1950 年到现在），我们正同样经历着一个相似的飞跃——在获取信息的能力上的飞跃

第 2 章

获取序列数据和文献信息

2.1 生物学数据库介绍

所有生命体所共有的一个特征是繁殖和进化的能力。一个物种的基因组被定义成该物种中所有的 DNA 的集合，包括那些编码蛋白质的基因。1995 年，人类测定了第一个能独立存活的生命体的基因组，这个生命体就是一种名叫 *Haemophilus influenzae* 的细菌（Fleischmann 等，1995；见第 12 章和第 14 章）。在这之后的若干年，更多物种的完整的基因组被测定出来，人类从此进入了一个生物学研究的新时代，这个时代的标志是生物学数据的高速积累以及各种生物学信息的方便地获取。目前，公共的数据库中已经保存了来自于上千个不同物种的数十亿个碱基的 DNA 序列数据。本章的目的是介绍储存这些数据的数据库，以及从这些数据库中获取需要的信息的方法。

世界上有三个大型的公共数据库存储着大量的核酸和蛋白质序列数据（表 1-1），它们分别是位于 Bethesda 的美国国家生物技术信息中心（NCBI）的 GenBank，日本的 DNA 数据库（DDBJ），以及由位于英格兰 Hinxton 的欧洲生物信息研究所（EBI）所维护的数据库。我们将在本章中介绍这 3 个数据库。

这 3 个数据库作为国际核酸序列数据库合作组织的成员，每天都相互共享更新数据。

除了 GenBank、DDBJ 和 EBI 之外，还有其他类型的存储 DNA 和/或蛋白质序列数据的生物信息数据库：

① 提供人类基因组和其他基因组注释的数据库，如 Ensembl、NCBI 以及位于加州大学圣克鲁斯分校（UCSC）的 Golden Path 数据库（见第 16～第 17 章）。

② 存储特定的基因或者蛋白质（如激酶）的核酸和/或蛋白质序列数据库。另外还有一些数据库记录的对象是特定的染色体或者细胞器的数据（见第 16～第 18 章）。

③ 有一系列数据库记录了共享了某种性质的序列的信息。例如蛋白质家族数据库（Pfam）就存储了几千个同源蛋白质家族的数据。

④ 有上百个数据库存储了和人类的疾病相关基因突变的序列信息。我们在第 18 章中介绍这些数据库。

⑤ 许多数据库是针对特别的物种（如酵母）而设计的；在有关基因组的章节（第 12～第 17 章）中列出了这类数据库的例子。

⑥ 一些数据库记录了特殊类型的核酸或蛋白质，或者和这些大分子有关的某种性质的数据，如基因表达谱数据库（见第6章和第7章），转运 RNA（tRNA）分子的数据库，组织特异性的蛋白质表达数据库（见第8章），或者存储基因调控区域（如3'端未转录区域）的数据库（见第16章）。

⑦ 有些数据库存储的核酸或蛋白质序列不是公开的。如 Celera Genomics 这样的公司提供了有偿的数据服务，必须购买使用权才能使用这些数据库（见第17章有关人类基因组的部分）。

有些数据库的主要功能并不是存储核酸或蛋白质序列数据，而是存储有关个别的基因/蛋白质的信息。

⑧ 文献数据库存储了生物学研究的参考信息和文献，部分还提供了全文的链接。本章我们将讨论其中的两个数据库：PubMed 和序列获取系统（SRS）。

Pfam（http://www.sanger.ac.uk/Software/Pfam）和其他相关数据库将在第8章中介绍。

⑨ 结构数据库记录了蛋白质和其他大分子的结构信息。我们将在第8章（蛋白质）和第9章（蛋白质结构）中讨论这些数据库。

2.2 GenBank：收录几乎所有已知的核酸和蛋白质序列的数据库

虽然在 DDBJ、EBI 以及 GenBank 中的序列信息是等同的，我们还是先从 GenBank 开始讨论。GenBank 数据库收录了目前绝大部分已知的公开的 DNA 和蛋白质序列（Benson 等，2002）。除了序列信息以外，GenBank 还存储了相应的参考文献记录以及生物学注释。GenBank 中的数据可以完全免费地从 NIH 的美国国家医学图书馆的美国国家生物技术信息中心（NCBI）获取（Benson 等，2002）。

第5章中，我们将介绍从各个基因组测序计划出发来搜索数据库的特殊方法。其中很多测序计划的序列数据是公开的，但还没有收录到 GenBank 中。

2.2.1 序列数据的总量

GenBank（第135.0版，2003年4月）收录了2400万条序列，总长度超过310亿个碱基。图2-1总结了从1982～2000年 GenBank 中 DNA 序列和长度的变化情况。从1982年到现在 GenBank 中序列的长度大概每14个月就翻一番。

2.2.2 GenBank 中的物种数

GenBank 中收录了超过10万个不同物种的数据，其中每1个月就会增加超过1000个新的物种（Benson 等，2002）。表2-1中显示了 GenBank 中各大类物种的数目。我们将在第14～第16章中对几个大类，也就是细菌、古细菌和真核生物进行详细地定义。简单地说真核生物都有一个细胞核而且常常是多细胞的，而细菌没有细胞核。古细菌是单细胞生物，它不同于真核生物和细菌，是生物界的第三大分支。病毒带有核酸（DNA 或 RNA），但是它只能在宿主的细胞中复制，在广义的生命形式的定义下，也被看成是生命的另外一个分支。

2001年10月～2002年10月，GenBank 中增加了约120亿个碱基对的 DNA 序列，平均每天增加3300万个碱基对。第一个完成测序的基因组长度约为1200万个碱基对（*Saccharomyces cerevisiae* 的基因组；见第15章）。

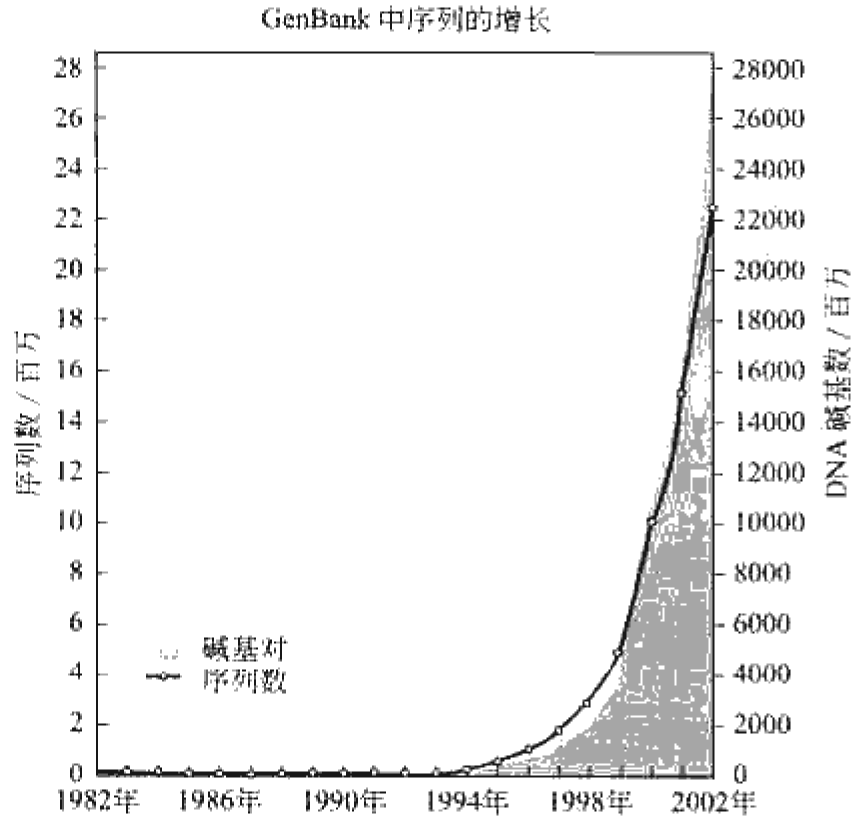


图 2-1 GenBank 中序列的增长（http://www.ncbi.nlm.nih.gov/GenBank/genbankstats.html）

表 2-1 GenBank（1995～2001）中收录的物种

项　目	1995 年	1996 年	1997 年	1998 年	1999 年	2000 年	2001 年
所有物种/个	15925	22862	32569	43159	61525	87168	113940
病毒/个	1740	1964	2513	2794	3401	4165	5436
细菌/个	2899	3798	6015	8625	14209	22616	28186
古细菌/个	154	227	376	544	1003	1697	2038
真核生物/个	10339	15863	22539	29844	41295	56800	76099

注：来源于 http://www.ncbi.nlm.nih.gov/Taxonomy。

我们已经看到，GenBank 数据库非常庞大，而且其数据量增长非常迅速。从表 2-1 我们可以看到 GenBank 中收录的物种大部分是真核生物。对于微生物来讲，目前 GenBank 中收录的细菌的物种数是其收录的古细菌的物种数的 10 倍以上。

你可以从 ftp://ftp.ncbi.nih.gov/genbank 上下载 GenBank 中所有的序列数据。2002 年 10 月发布的第 132.0 版的 GenBank 中，所有数据文件总的大小约为 83G（83×10^{9} 比特）。作为比较，美国国会图书馆的所有馆藏文字加起来总共有 20T 比特（20×10^{12} 比特，20 万亿比特）。位于日内瓦附近的 CERN（http://public.web.cern.ch/Public/）的物理学家们使用粒子加速器每年能收集 10^{15} 比特（即 1000 万亿比特）的数据。

表 2-2 总结了 GenBank 中 20 种测序最多的物种的 DNA/RNA 序列的条数和总长度（不包括叶绿体和线粒体的序列）。这个列表中包括了几种在生物学中研究得最多的模式生物。值得注意的是，科学界正在研究一系列哺乳动物（如人、鼠和牛等）、无脊椎动物（果蝇、线虫）、植物（洋蓟、谷物、水稻）以及寄生生物（非洲锥虫、疟原虫）。不同的物种对许多不同的研究课题来说有着各自的用途。细菌、古细菌以及病毒没有在表 2-2 中给出，这是因为它们的基因组都相对较小的缘故。

为了更好地组织数据，GenBank 中记录的每条序列的名称后都紧跟着它的数据文件类型以及主要的索引号码（我们将在下文中给出索引号码的定义）。GenBank 用以下的代码来表示数据文件的类型。

表 2-2　GenBank 中 20 种测序最多的物种（2003 年 4 月发布的第 135.0 版）

DNA/RNA 序列的条目/条	碱基/个	物　种	通 用 名
6574171	9743398611	*Homo sapiens*	人
4839873	5815119777	*Mus musculus*	小鼠
702180	5565107808	*Rattus norvegicus*	大鼠
481072	725258089	*Danio rerio*	斑马鱼
350741	695203285	*Drosophila melanogaster*	果蝇
221048	620437677	*Oryza sativa (japonica cultivar-group)*	水稻
730719	402816176	*Zea mays*	玉米
502590	394829161	*Arabidopsis thaliana*	拟南芥
567863	386355126	*Brassica oleracea*	甘蓝
13379	326808471	*Macaca mulatta*	恒河猴
427323	317823422	*Gallus gallus*	鸡
499207	294135749	*Ciona intestinalis*	海鞘
390231	223252029	*Bos taurus*	牛
199302	222094921	*Caenorhabditis elegans*	蠕虫
418485	207706076	*Triticum aestivum*	小麦
166243	191992299	*Pan troglodytes*	黑猩猩
189208	170811939	*Tetraodon nigroviridis*	河豚鱼
273621	166171916	*Xenopus laevis*	爪蟾
282062	156196664	*Hordeum vulgare subsp. vulgare*	大麦
186121	154322129	*Medicago truncatula*	蒺藜状苜蓿

注：来源于 ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt。

① PRI：灵长类（primate）的序列。

② ROD：啮齿类（rodent）的序列。

③ MAM：其他哺乳类（mammalian）的序列。

④ VRT：其他脊椎动物（vertebrate）的序列。

⑤ INV：非脊椎动物（invertebrate）的序列。

⑥ PLN：植物、真菌和藻类的序列。

⑦ BCT：细菌（bacterial）的序列。

⑧ VRL：病毒（viral）的序列。

⑨ PHG：噬菌体（bacteriophage）的序列。

⑩ SYN：人工合成（synthetic）的序列。

⑪ UNA：未经注释（unannotated）的序列。

⑫ EST：EST 序列（expressed sequence tags，表达序列标签）。

⑬ PAT：专利（patent）序列。

⑭ STS：STS 序列（sequence-tagged sites，序列标签位点）。

⑮ GSS：GSS 序列（genome survey sequences，基因组测序序列）。

⑯ HTG：HTGS 序列（high-throughput genomic sequences，高通量基因组序列）。

我们将在第 12 章讨论测序的物种是如何选取的。

国际人类基因组测序委员会在 1996 年通过了百慕大协议（the Bermuda Priciples），指出要快速发布基因组序列的原始数据。你可以从网站上阅读到这个协议的最新版本：http://www.genome.gov/page.cfm?pageID=10506537。

2.2.3 GenBank 存储的数据类型

GenBank 中的分子序列数据量是很大的。我们在下文中将介绍 GenBank 中基本的数据类型，然后将讨论如何从 GenBank 中获取你需要的数据。

我们从一个例子开始。假定我们要获取人类视黄醇结合蛋白（retinol-binding protein，RBP4）的序列。要记住的重要的一点是 DNA 和蛋白质序列是存储在不同的数据库中的。而且在不同的数据库中，数据的表示形式也多种多样。例如，*RBP4* 可以在 DNA 层面上进行描述（作为一个基因）；或者在 RNA 层面上进行描述［为一个信使 RNA（mRNA）］；或者在蛋白质层面上进行描述（见图 2-2）。因为 RNA 相对来说比较不稳定，通常把 RNA 转换成相应的互补 DNA（cDNA），有一系列的数据库收录了这样的 cDNA 数据，每个 cDNA 都对应一个 RNA 转录片段。这样对我们使用例子 *RBP4* 来说，有如下几种数据形式。

我们将在下文中给出 STS、GSS、EST 和 HTGS 的具体定义。

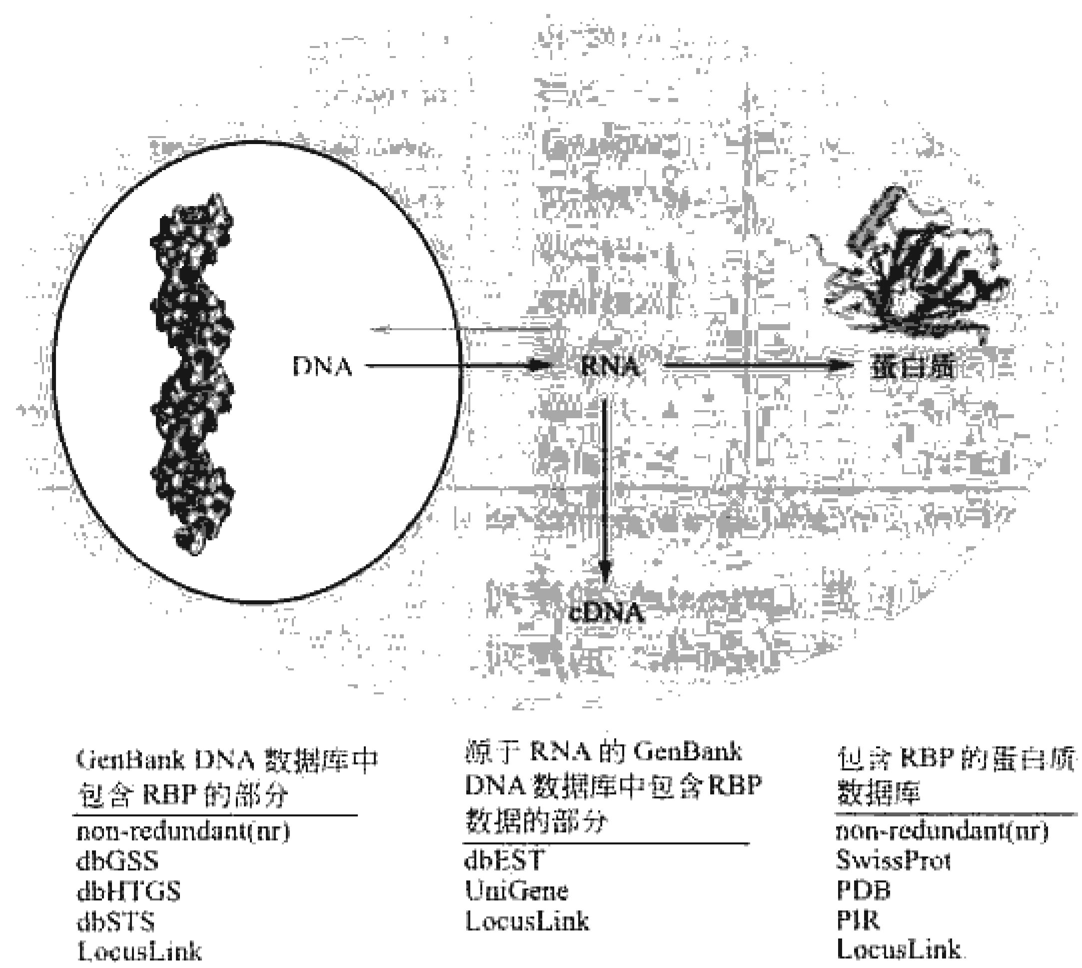


图 2-2 GenBank 和其他数据库中的序列数据类型（以 RBP4 为例）。注意到"retinol-binding protein"可能指的是一个基因、一个 RNA 转录片段（或其对应的互补 DNA），或者一个蛋白质。这三种分子都有各自对应的专门的数据库。当然还有许多数据库不在 GenBank 和 NCBI 范围内（没有列出）；注意到 SwissProt、PDB 以及 PIR 数据库都是独立于 GenBank 的蛋白质数据库。GenBank、DDBJ 和 EBI 中的核酸序列是相同的

2.2.3.1 基因组 DNA 数据库

① *RBP4* 是染色体的一部分。以人类的 *RBP* 为例，我们可以看到这个基因位于第 10 号染色体上（第 17 章，有关人类基因组的部分）。

RBP4 也被称为 RBP。通常基因并不和它所编码的蛋白共用同一个名字。HD 是引起舞蹈症（Huntingon disease）的基因的名称，但是它所编码的产物（蛋白质）则被称作 huntingtin。不同的研究人员在研究相同的基因的时候常常会给这个基因起不同的名称。人类基因组组织（HGUGO）的基因命名委员会（HGNC）的任务非常重要，它必须给基因和蛋白质提供官方的命名（见 http://www.gene.ucl.ac.uk/nomenclature/）。

② *RBP4* 作为一些可能带有多个基因的大片段 DNA 的一部分，如黏粒、细菌人工染色体（BAC）或者酵母人工染色体（YAC）的一部分。BAC 是克隆到细菌中的大片段 DNA（长度通常是 20 万个碱基对，或者 200kb）。类似的 YAC 用于在酵母中克隆大片段的 DNA

序列。BAC 和 YAC 在对基因组的大片段进行测序的时候非常有用。

③ *RBP4* 作为一个基因存储在数据库中。这个基因是遗传的功能单位，同时它也是一个 DNA 序列，通常还带有调控区；其编码蛋白质的外显子和内含子。人类的基因长度通常是 10～50kb。

第 10 号染色体是一个中等大小的染色体，它约含有 1500 个基因，长度约为 15 万 kb。

④ *RBP4* 作为一个序列标签位点（STS）——也就是作为用来连接基因图谱和物理图谱的小 DNA 片段（通常 500bp），在序列标签位点的数据库（database of sequence-tagged sites，dbSTS）中存储。

2.2.3.2 对应于表达基因的 cDNA 数据库

RBP4 作为一个表达序列标签（EST）存储在数据库中，也就是作为一个从特定的 cDNA 库中得到的一段 cDNA 序列。如果我们得到了一个组织（如肝脏），从中我们提取 RNA，然后把 RNA 转化成更稳定的 cDNA，一些这样的 cDNA 库中的某些 cDNA 克隆很可能就编码了 *RBP4*。

2.2.3.3 蛋白质数据库

RBP4 这个名称可能指 DNA、RNA 或者蛋白质。作为一个蛋白质，RBP4 被收录在一系列数据库中，如 GenBank 的非冗余数据库 nonredundant（nr）database、SwissProt 数据库（Boeckmann 等，2003）以及蛋白质数据库 Protein Data Bank（Westbrook 等，2002）中。

GenBank 的惯例是使用 DNA 序列来表示 RNA 分子。

2.2.4 表达序列标签（ESTs）

表达序列标签数据库是 GenBank 数据库的子库，它收录了一系列物种中的“只测序一次”的 cDNA 序列数据和相关信息（Boguski 等，1993）。一个 EST 是一个 cDNA 克隆的一部分序列。所有的 cDNA 克隆，因而所有的 EST 数据都来源于特定的 RNA 源，如人脑或者大鼠的肝脏。RNA 被转换成更稳定的 cDNA 分子，许多 cDNA 分子组成一个 cDNA 文库（见图 2-2）。EST 序列通常是随机从 cDNA 克隆中选出，然后对一条链进行快速地测序。EST 的长度通常是 300～800bp。最早的 EST 测序工作发现了几百个在当时还未知的基因（Adams 等，1991）。

2003 年 5 月，GenBank 保存了超过 1600 万条的 EST 记录，其中包括了在表 2-3 中列举的“前 10 位”的物种的数据。如果我们假定人类基因的总数是 35000 个，那么平均每个基因将对应 150 个 EST 记录。

目前 GenBank 把 EST 分成三个大类：人、鼠和其他生物。

表 2-3 EST 已被测序的前十个物种（dbEST 第 050903 发布版本，2003 年 5 月）

物种	通用名	EST 条目的数目/条
Homo sapiens	人	5142390
Mus musculus＋*domesticus*	小鼠	3721428
Rattus sp.	大鼠	525556
Ciona intestinalis	海鞘	492488
Gallus gallus	鸡	418093
Triticum aestivum	小麦	340945
Hordeum vulgare＋subsp.	大麦	340945
Bos taurus	牛	319775
Danio rerio	斑马鱼	311335
Glycine max	大豆	308582

注：1. 从各个物种中得到的 cDNA 文库有好几千个，数据库中的公共条目的数目大概是 1600 万。

2. 来源于 http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html。

ESTs 和 UniGene。UniGene（unique gene）项目的目标是为每一个基因创造一个惟一的条目，然后收集对应于这个基因的所有 EST。以 *RBP4* 为例子，对应于 *RBP4* 只有一个 UniGene 条目。表 2-4 给出了 UniGene 中收录的 26 个物种。

表 2-4 UniGene 中收录的物种

属	物 种	缩 写	通 用 名
脊索动物门	*Bos taurus*	Bt	牛
	Ciona intestinalis	Cin	海鞘
	Danio rerio	Dr	斑马鱼
	Gallus gallus	Gga	鸡
	Homo sapiens	Hs	人
	Mus musculus	Mm	小鼠
	Oryzias latipes	Ola	日本鲫
	Rattus norvegicus	Rn	大鼠
	Silurana tropicalis	Str	西方爪蟾
	Sus scrofa	Ssc	猪
	Xenopus laevis	Xl	非洲爪蟾
节肢动物门	*Anopheles gambiae*	Aga	疟蚊
	Drosophila melanogaster	Dm	果蝇
蠕虫动物门	*Caenorhabditis elegans*	Cel	蠕虫
有胚植物	*Arabidopsis thaliana*	At	拟南芥
	Glycine max	Gma	大豆
	Hordeum vulgare	Hv	大麦
	Lycopersicon esculentum	Les	番茄
	Medicago truncatula	Mtr	蒺藜状苜蓿
	Oryza sativa	Os	水稻
	Solanum tuberosum	Stu	马铃薯
	Sorghum bicolor	Sbi	高粱
	Triticum aestivum	Ta	小麦
	Zea mays	Zm	玉米
绿藻门	*Chlamydomonas reinhardtii*	Cre	绿藻
黏菌门	*Dictyostelium discoideum*	Ddi	黏液菌

注：来源于 http://www.ncbi.nlm.nih.gov/UniGene/（2003 年 5 月）。

虽然 *RBP4* 只有一个 UniGene 条目，但是目前有超过 200 个人类的 EST 和 *RBP* 基因相对应。如此大量的 EST 数据提示出 *RBP* 在已经被测序的 cDNA 文库中表达量非常大。UniGene 簇是一个基因的数据库条目，这个数据库条目包含了所有和这个基因对应的 EST（图 2-3）。

为了搜索 RBP 相关的数据，我们链接 http://www.ncbi.nlm.nih.gov 的网页，在 Search 列表中选择 UniGene，再选择 "human"，然后输入 "retinol-binding protein" 或 RBP4，得到的 UniGene 索引号是 Hs50223（原版书中的 Hs.76461 已不再使用。——译者注）；注意这里 "Hs" 指的是 "*H. sapiens*"。

UniGene 簇是通过 EST 的序列相似性搜索，收集匹配的序列来创建的（见第 3 章和第 4 章）。许多簇仅仅只有一个成员，代表着独特的序列，而其他的基因簇却可能有几千个 EST 成员。我们将在第 6 章（关于基因表达）中更深入地讨论 UniGene 基因簇。

现在人们普遍认为人类的基因总数约为 35000 个（见第 17 章）。你可能觉得 UniGene 的基因簇也应该差不多有这么多。但是，实际上现在 UniGene 基因簇的个数比基因的个数多了许多，数据库中约有 10 万个人类的基因簇。这种差异可能来源于三个原因。①一个 ESTs 簇可能对应于一个基因的不同部分（见图 2-3）。随着基因组的测序变得更完全，这样的两个簇很可能会被聚成一个簇。

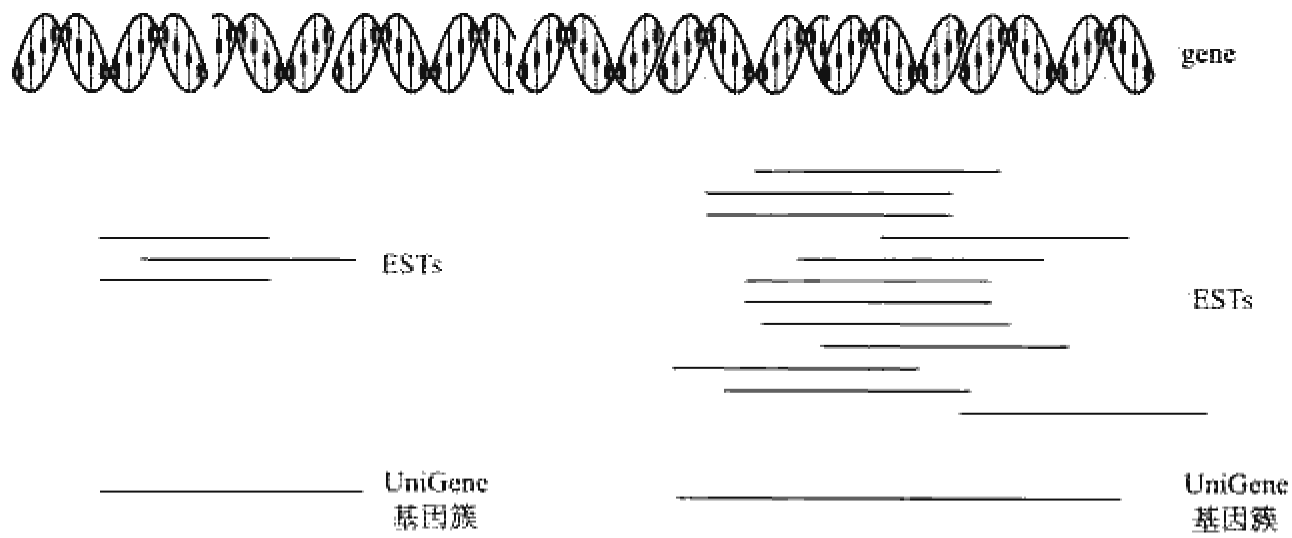


图 2-3 用图形方式表示 UniGene 簇。表达序列标签 EST 被映射到一个特定的基因上，同时 EST 之间也有对应关系。对应于一个 UniGene 簇的 EST 的个数范围是 1～1000；平均来说一个簇对应约 100 个 EST 序列。有时候，不同的簇对应于同一个基因的不同部分，如图中所示，这两个簇最终将变成一个簇。最后，UniGene 簇的个数将等同于基因组中基因的个数

我们使用 RBP4 作为例子。如果你输入的是"retinol-binding protein"，不论你访问哪一个数据库你都能得到更多的结果。而在 UniGene 中，你将搜索到几十个条目。

这样 UniGene 簇的个数可能随着时间的推移逐渐变少。②有些基因表达水平可能特别低。35000 个人类的 UniGene 簇只含有一个 EST，而超过 6 万个 UniGene 簇仅含有 1～4 个 EST 记录。这些可能反映了一些还无法用其他基因识别方法识别出来的真实的基因。③一些 DNA 可能在建立 cDNA 库的时候被转录出来，但是并不对应任何一个基因，可能是克隆过程中的副产品。我们将在第 15～第 17 章中讨论有哪些特征定义了一个真核基因。可变剪切（第 6 章）可能会为一个基因带来不同的簇，因为被剪切掉的外显子和序列的其他部分并没有同源性。

UniGene 簇条目数量引人注目的上升如图 2-4 所示。许多 UniGene 簇条目是源自 CGAP（癌基因解剖学计划），关于这个计划的介绍可参见 http://cgap.nci.nih.gov/。

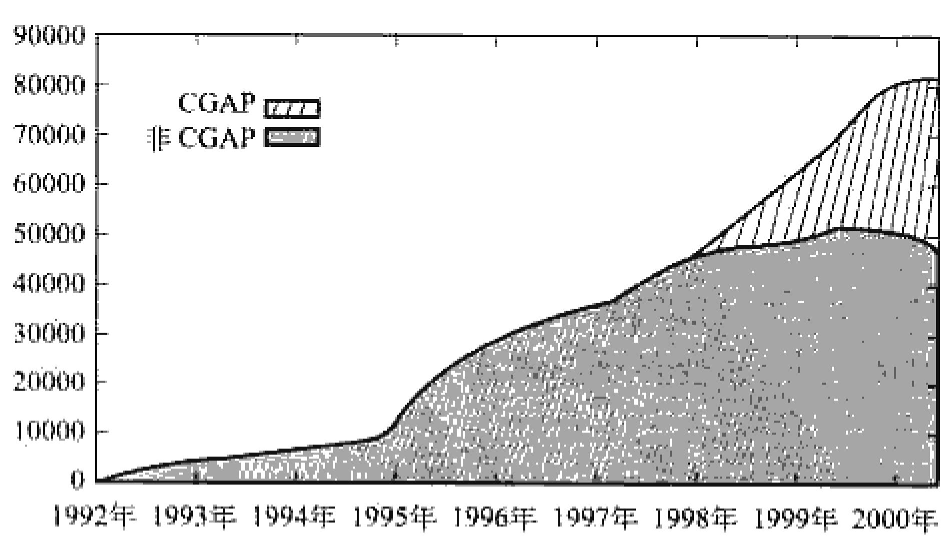


图 2-4 UniGene 中人类基因的记录数的增长情况（http://www.ncbi.nlm.nih.gov/UniGene/）。有关 Cancer Gene Anatomy Project（CGAP）的介绍可从 http://cgap.nci.nih.gov/获得

2.2.5 序列标签位点（STS）

dbSTS 是 NCBI 中记录 STS 的数据库，STS 是基因组中较短的用于定位的序列，STS

2003年2月，数据库中保存有来自几个常用的模式生物的226000条STS序列记录（表2-5）。

记录包括DNA序列数据和图谱数据（Olson等，1989）。约有十几个物种的STS数据已经被测定出来（表2-5）。一个通常的STS序列的尺寸大约和EST序列相当。STS常常具有多态性，并且带有短的重复序列（第16章），因此在进行图谱研究的时候STS有非常重要的作用。

2.2.6 基因组测序序列（GSSs）

目前，数据库中共有约4600万个GSS数据条目（其中约有一半数据来自人类和啮齿类动物）（2003年2月）。GSS数据库的网站是http://www.ncbi.nlm.nih.gov/dbGSS/index.html。

GenBank中的GSS子项类似于EST子项，只不过它的序列来源于基因组，而非cDNA（mRNA）。GSS子项收录有如下几个类型的数据（见第16章和第17章）：

① 随机的"经过单次测序"的基因组测序序列；

② 黏粒/BAC/YAC序列；

③ 外显子定位的基因组序列；

④ Alu聚合酶链反应（PCR）序列。

表2-6给出了具有GSS数据的物种。

表2-5 具有STS数据的物种

物种名称	STS序列的大概数量/个
人(*Homo sapiens*)	134000
挪威大鼠(*Rattus norvegicus*)	30000
家鼠(*Mus musculus*)	27000
Danio rerio	22000
Drosophila melanogaster	1100

注：来源于http://www.ncbi.nlm.nih.gov/genome/sts/unists.stats.html（2000年11月）。

表2-6 具有GSS数据的物种

物种名称	GSS序列的大概数量/个
家鼠(*Mus musculus*)	945000
人(*Homo sapiens*)	873000
甘蓝(*Brassica oleracea*)	567000
挪威大鼠(*Rattus norvegicus*)	307000
拟南芥(*Arabidopsis theliana*)	214000
河豚鱼(*Tetraodon nigroviridis*)	189000

注：来源于http://www.ncbi.nlm.nih.gov/dbGSS/dbGSS.summary.html（2003年2月）。

2.2.7 高通量基因组序列（HTGS）

为了让科学界更快地得到目前还没有完成的基因组序列数据，人们建立了HTGS数据库。它是由三个跨国的核算序列数据库（DDBJ、EMBL和GenBank）合作建立的。HTGS数据库记录了由高通量的测序中心测出的尚未完工的DNA序列。HTGS的主页地址是http://www.ncbi.nlm.nih.gov/HTGS/，我们可以用BLAST（见第4章和第5章）来搜索其中的序列。

我们已经讨论了GenBank中一些基本的序列数据类型。下面我们将转而讨论Entrez和位于NCBI以及别的地方的其他一些项目，这些项目可以让你获取GenBank、EMBL和DDBJ中的数据和相关的文献信息。我们介绍NCBI网站，这是生物信息学领域最重要的一个基于互联网的资源。

2.3 美国国家生物技术信息中心（NCBI）

2.3.1 NCBI的介绍：NCBI的主页

NCBI的主要工作是开发数据库，进行计算生物学研究，开发用于分析基因组数据的软

件工具，发布生物医学信息（Wheeler 等，2002）。图 2-5 给出了 NCBI 的主页。网站上方的导航条有七个大类：PubMed、Entrez、BLAST、OMIM、Books、Taxonomy 和 Structure。

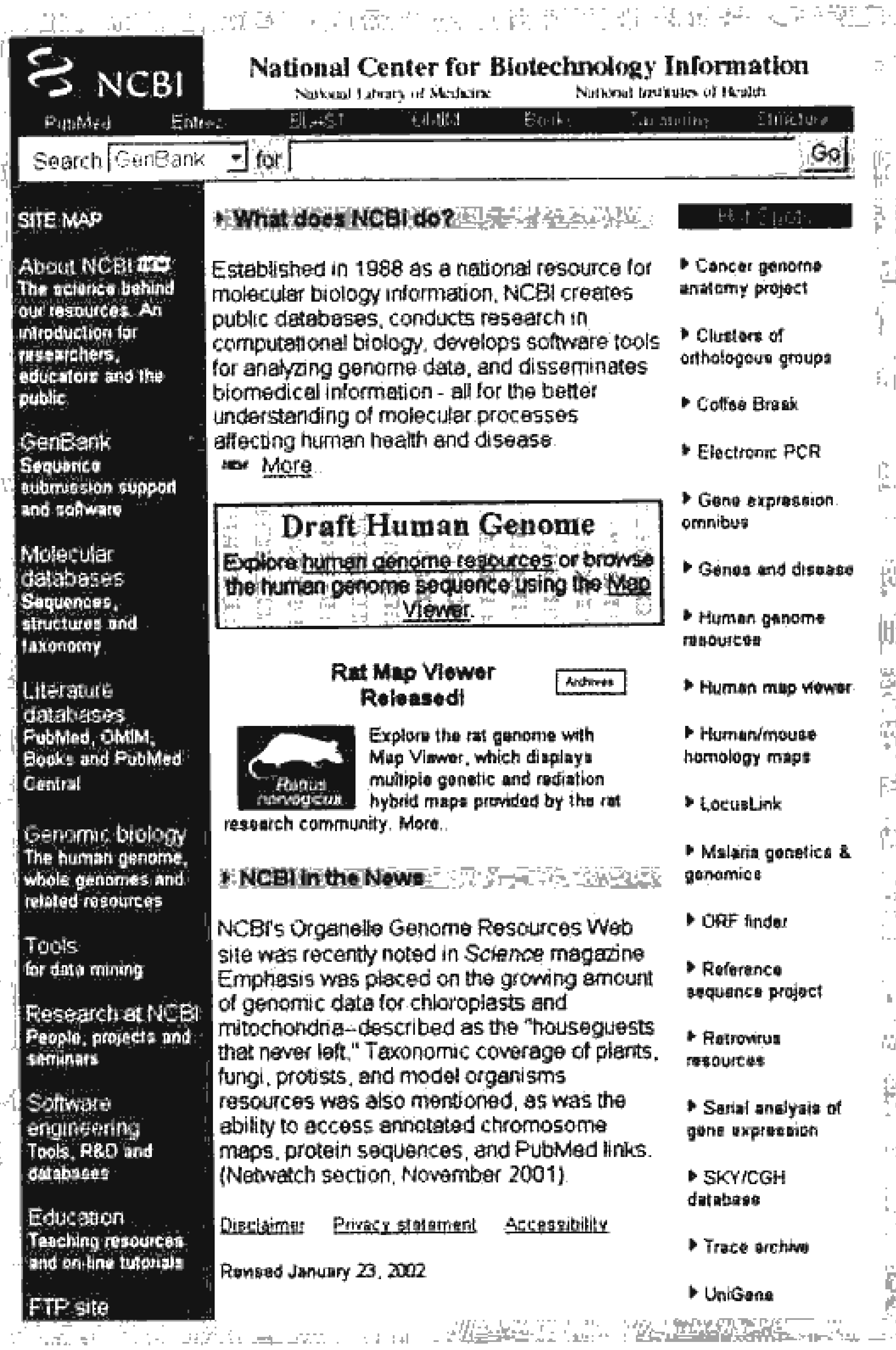


图 2-5 美国国家生物技术信息中心（NCBI）网站主页（http://www.ncbi.nlm.nih.gov）。导航栏上的子项目包括 PubMed、Entrez and Books（本章中介绍）、BLAST（见第 3～第 5 章）、Taxonomy（第 12～第 17 章）、Structure（第 9 章）和 Online Mendelian Inheritance in Man（OMIM，第 18 章）。请读者留意左边一栏的教育（Education）子项中还列出了若干教程

2.3.1.1 PubMed

PubMed 是美国国家医学图书馆（NLM）提供的搜索服务，它提供了来自 MEDLINE（医学文献在线分析、获取系统）和其他相关数据库的超过 1100 万条的文献记录，同时还提供了到许多在线期刊的链接。NCBI 的网页还提供了一个 PubMed 教程，点击 NCBI 主页（http://www.ncbi.nlm.nih.gov）中左边第一列的“Education”，然后点击 PubMed 教程的链接。

2.3.1.2 Entrez

Entrez 将科学文献、DNA 和蛋白质序列数据库、蛋白质三维结构数据、种群研究数据以及全基因组组装数据整合成一个高度集成的系统。PubMed 是 Entrez 中专门针对文献的部分。

2.3.1.3 BLAST

BLAST（basic local alignment search tool，基础局部比对搜索工具）是为了分析核酸和蛋白质数据库而设计的序列相似性搜索工具（Altschul等，1990，1997）。BLAST由一套相似性搜索工具组成，这套搜索工具可以搜索所有序列数据库，无论针对的是蛋白质还是DNA序列。我们将在第3～第5章中讨论BLAST。

2.3.1.4 OMIM

在线人类孟德尔遗传性状数据库（online mendelian inheritance in man，OMIM）是一个人类基因和遗传异常的索引。Victor McKusick和他的同事们发明了这个数据库，然后NCBI针对国际互联网对数据库进行了进一步开发（Hamosh等，2002）。这个数据库包含了详细的参考信息，其中还有与PubMed文章和序列信息的链接。我们将在第18章中讨论OMIM（有关人类疾病的信息）。

2.3.1.5 Books

NCBI提供了十多部在线的参考书籍。这些书籍都是可搜索的，并且包含了与PubMed的链接。

2.3.1.6 Taxonomy

NCBI的Taxomony网站包含了一个囊括了主要的生物类别的分类信息浏览器（古细菌、细菌、真核生物和病毒）。网站还提供了一系列分类信息，例如遗传编码和分类资源以及其他附加的信息（如灭绝物种的分子信息和分类系统最近改变的信息）。我们将在第11章中（关于进化方面）和第12章和第17章（关于基因组和生命分类树方面）中访问这个站点。

> 蛋白质数据库（Protein Data Bank，PDB）（http://www.rcsb.org/pdb）是互联网上惟一的一个处理和发布生物大分子结构数据的网站。我们将在第9章介绍PDB。

2.3.1.7 Structure

NCBI的Structure结构网站提供了分子建模数据库（MMDB），这是一个记录了大分子三维结构，以及使这些结构可视化和进行比较分析的工具。MMDB包含了蛋白质数据库（protein data bank，PDB）中存储的实验测定的生物大分子的结构数据。NCBI提供的结构方面的资源还包括PDBeast（MMDB提供的分类信息）、Cn3D（一个三维结构可视化工具），以及一个允许对结构进行比较的向量队列搜索工具（VAST）（见第9章有关蛋白质结构的部分）。

2.3.2 索引号码：识别序列的标签

当你在研究基因或蛋白质遇到了问题时，你很可能需要从一些数据库的记录中寻找需要的信息。你可能要先从文献中获取信息或者你必须有感兴趣的序列的名字，从这些信息开始解决你要研究的问题。也许你有一段氨基酸序列和/或核算序列数据；我们将在第3～第5章中讨论这些内容。现在我们要讨论的是如何从数据库中获取你感兴趣的基因或蛋白质信息。

> DNA测序时通常是两条链同时测序，但是EST在测序时常常只有一条链，这导致了EST较高的错误率。我们将在第12章讨论测序的错误率。

DNA和蛋白质序列的一个重要的特征是它们都被打上了索引号码作为标签。索引号码是一段由约4～10个数字和/或字符组成的编码，每个索引号码与一个分子的序列记录相对应。索引号码也可能用来作为其他的数据记录的标签（如蛋白质结构），甚至一个基因表达实验的结果（第6章和第7章）。标识不同数据库的记录索引号码有各自不同的格式特点（框2-1）。导致这些格式不同的原

因在于各自的数据库采用了不同的系统。当你开始探索这些数据库，从中提取 DNA 和蛋白质数据的时候，你要尽快地熟悉这些索引号码的不同的格式。有些数据库（图 2-2）的索引号可以告诉我们它所对应的序列属于哪一类型（是核酸序列还是蛋白质序列）。

框 2-1 索引号码的类型

记录类型	索引号码格式的例子
GenBank/EMBL/DDBJ 核酸序列记录	一个字母跟着 5 个数字，如 X02775 两个字母跟着 6 个数字，如 AF025334
GenPept 序列记录（从 GenBank/EMBL/DDBJ 的记录中带有编码区注释的序列翻译过来的氨基酸序列数据库）	三个字母跟着 5 个数字，如 AAA12345
SwissProt 和 PIR 的蛋白质序列记录	通常是一个字母跟着 5 个数字，如 P12345。SwissProt 索引号码也可能是数字和字母的混合排列
蛋白质研究基金会提供的蛋白质序列记录	一系列的数字（通常是 6～7 个数字）跟着一个字母，如 1901178A
RefSeq 核酸序列记录	两个字母，1 个下划线和 6 个数字，如 mRNA 记录（NM_*）：NM_006744；基因组 DNA 片段记录（NT_*）：NT008769
蛋白质结构记录	PDB 索引号码通常是 1 个数字跟着 3 个字母，如 1TUP。也可能是其他的数字和字母的混合排列（或者只由数字组成）。MMDB 的索引号码通常是 4 位数字，如 3973

注：本表修改自 http://www.ncbi.nlm.nih.gov/entrez/query/static/help/helpdoc.html。

以一个特定的分子——RBP4 为例，这个分子对应了几百个索引号码（图 2-6），其中很大部分对应于 EST 数据和其他一些和 RBP4 相匹配的 DNA 片段。你如何来衡量这些序列或蛋白质数据的质量呢？许多序列记录都带有一些错误，尤其是在 EST 记录的末尾处。

我们比较从 mRNA 推导出来的 RBP4 序列和从基因组 DNA 中推导出来的 RBP4 的序列，我们预计它们应该能完美地（或者至少是接近完美地）匹配。但是我们会看到，两条序列有很多地方都不一样（图 6-4）。最近在序列管理中一项重要的工作是对参考序列的开发，我们将在下文中对其进行介绍。RefSeq 的目标是为每一个基因正常的（也就是没有突变的）转录产物和每一个正常的蛋白质产物提供最好的代表序列（Maglott 等，2000）。

GenBank 或 RefSeq 的索引号指向序列的最新版本，例如 NM_006744.2 是目前人类 *RBP4* 对应的 RefSeq 记录，后缀 .2 是记录的版本号。默认情况下，不需要指定版本号，数据库会自动提供最新版本的序列。请试着用 NM_006744.1 在数据库中进行搜索，可以看到这个索引号码对应的记录的更新情况。

2.3.3 五种获取 DNA 和蛋白质序列的方法

我们如何才能在这么多数据库保存的大量的 DNA 和蛋白质序列数据中不至于迷失方向呢？这里我们将简要地介绍五个工具：LocusLink（包括重要的参考序列）、UniGene、Entrez、Ensembl 和 ExPASy。一个共同的背景是越来越多的数据库互相连接了起来，提供非常方便的交互链接以及一些用于分析 DNA、RNA 和蛋白质算法的链接。

2.3.3.1 LocusLink：基因和蛋白质信息的概括性资源

LocusLink 是一个辅助的数据库，它提供了有关遗传基因座的描述性信息（Wheeler 等，2002；Pruitt 等，2000）。你可以从 LocusLink 提供的信息中找到物种的官方命名、别名、序列索引号、表现型、EC 编号、MIM 编号、UniGene 簇、同源基因、图谱定位和相关网站等。

LocusLink 可以从 NCBI 的主页中进入，也可以直接通过 http://www.ncbi.nlm.nih.gov/LocusLink/ 访问。我们将在后面的章节中介绍更多 LocusLink 提供的资源，如 Enzyme Commission（EC）编号（第 8 章）和人类孟德尔遗传性状（NIM，见第 18 章）。

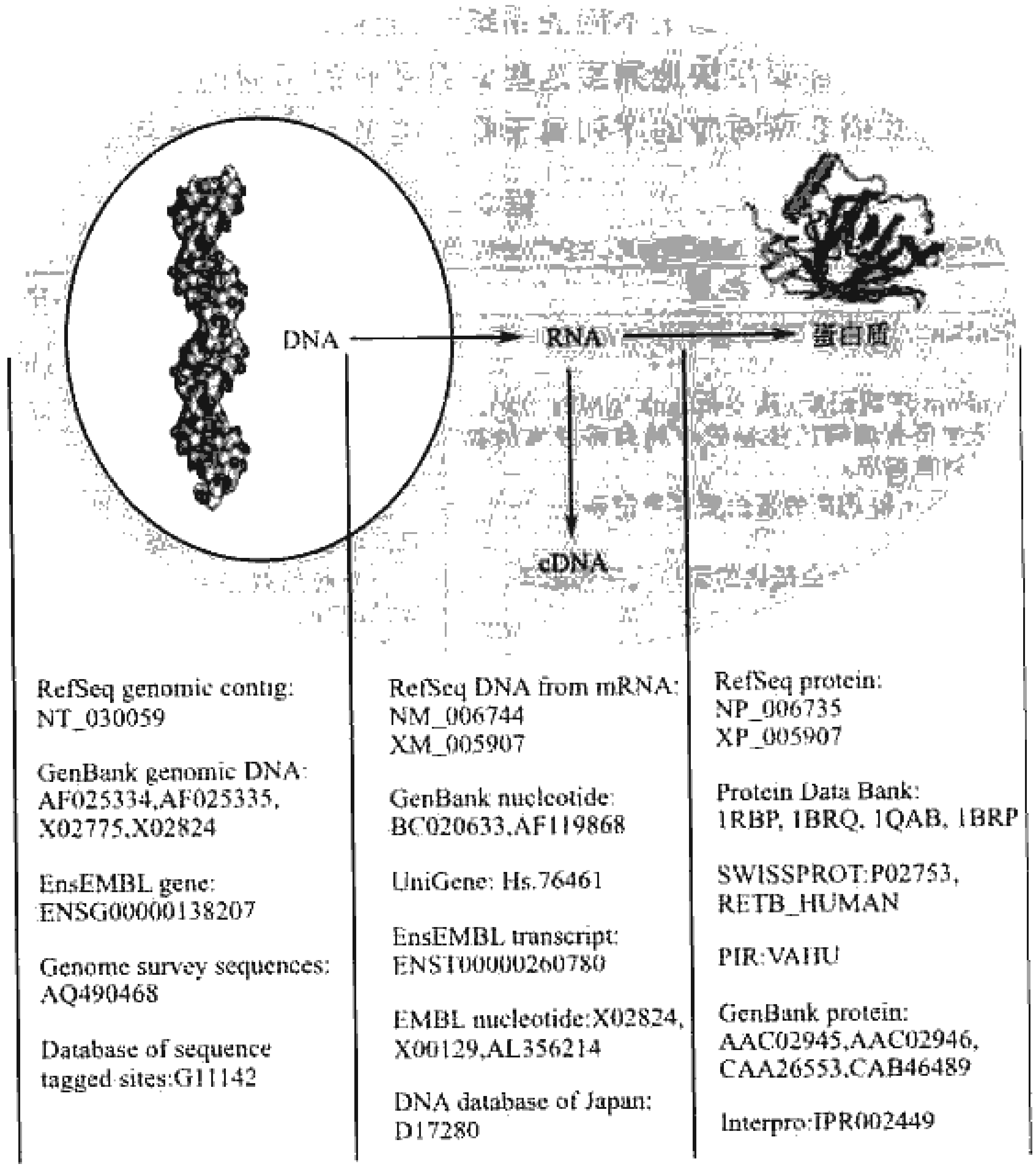


图 2-6 索引号码的例子（以 RBP4 为例）。索引号码以标签的形式指向从基因组 DNA 中得到的序列记录（图的左半部）；从 mRNA 得到的 cDNA 序列（包括 cDNA 测序得到的表达序列标签数据，图的中间部分）；以及蛋白质数据（图的最右边）。RefSeq 项目是非常重要的工作，它的目标是为每一个基因正常的（即没有突变的）转录产物和每一个正常的蛋白质产物提供最好的代表序列。其他的索引号码包括 LocusLink 的 LocusID（5950）以及 OMIM 的识别号码 180250。图中显示的数据库资源在第 2 章、第 6 章和第 8 章中讨论

LocusLink 为几个物种［人、小鼠、大鼠、果蝇（*Drosophila*）、斑马鱼、牛、线虫和 HIV-1］提供了快速、直观地获取蛋白质和 DNA 序列索引号的途径。特别需要注意的是 LocusLink 记录还为每一个 DNA 或者蛋白质提供了 RefSeq 的链接（见下文）。LocusLink 还提供了到 OMIM、HomoloGene（记录人、大鼠和小鼠中同源基因的数据库）、归类的图谱信息、基因组数据库和其他类似 GeneCard（有关基因主要功能的单页的概括性信息）的数据库的链接。

图 2-7 显示了用 “retinol-binding protein” 作为关键字在 LocusLink 搜索的结果。我们注意到进行 LocusLink 搜索的时候，可以方便地选择 “brief” 或 “summary” 来改变默认显示设置，也可以方便地把搜索限制在某一个特定的物种中。点击 LocusID 5950，这是与 main serum retinol-binding protein RBP4 的链接，得到如下的结果（见图 2-8）。

① 上面的导航栏给出了到各个数据库（如 PubMed、OMIM、UniGene 和 NCBI）图谱浏览器和描述单核苷酸多态性（SNPs，见第 18 章）的数据库中 RBP4 的链接。

② 搜索结果给出了 RBP4 功能的简要描述，定义 RBP4 的功能为一种称为 lipocalin 家族的转运蛋白。

③ 一个功能的子项目，它描述 RBP 缺陷的表现型。

④ 反映 RBP 基因在染色体中位置的图谱信息（10q23～24）。

⑤ 搜索结果还提供了参考序列（reference sequence，RefSeq）索引号：表示 DNA 编码的 RBP 的记录的索引号 NM _ 006744 和表示 RBP4 蛋白质的记录 NP _ 006735（见下文）。

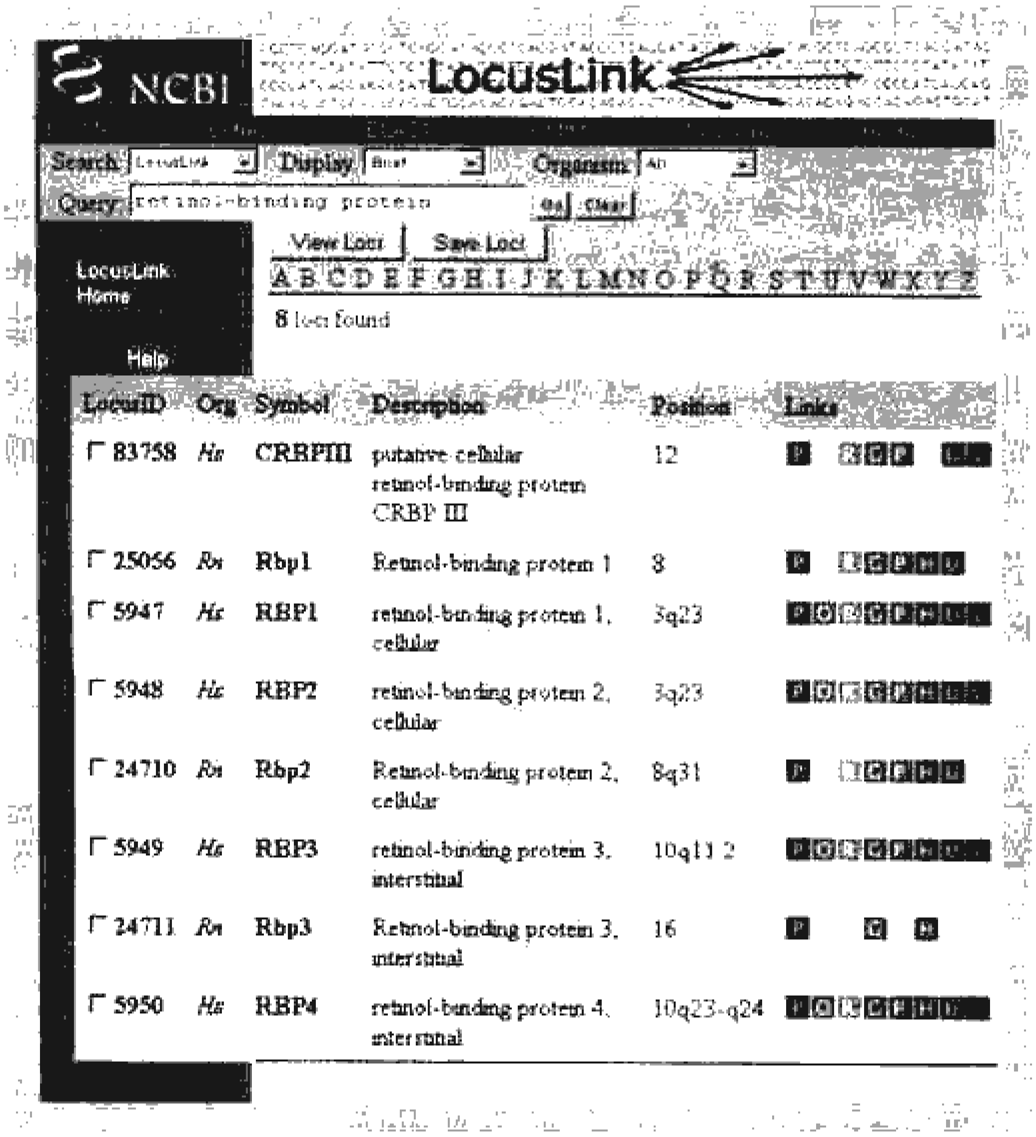


图 2-7　用"retinol-binding protein"在 LocusLink（http://www.ncbi.nlm.nih.gov/LocusLink/）中进行搜索的结果。结果中还给出了物种的信息，包括 *H. sapiens*（Hs）和 *R. norvegicus*（Rn），以及基因在染色体中的位置信息和其他一系列到不同数据库的链接（见正文）

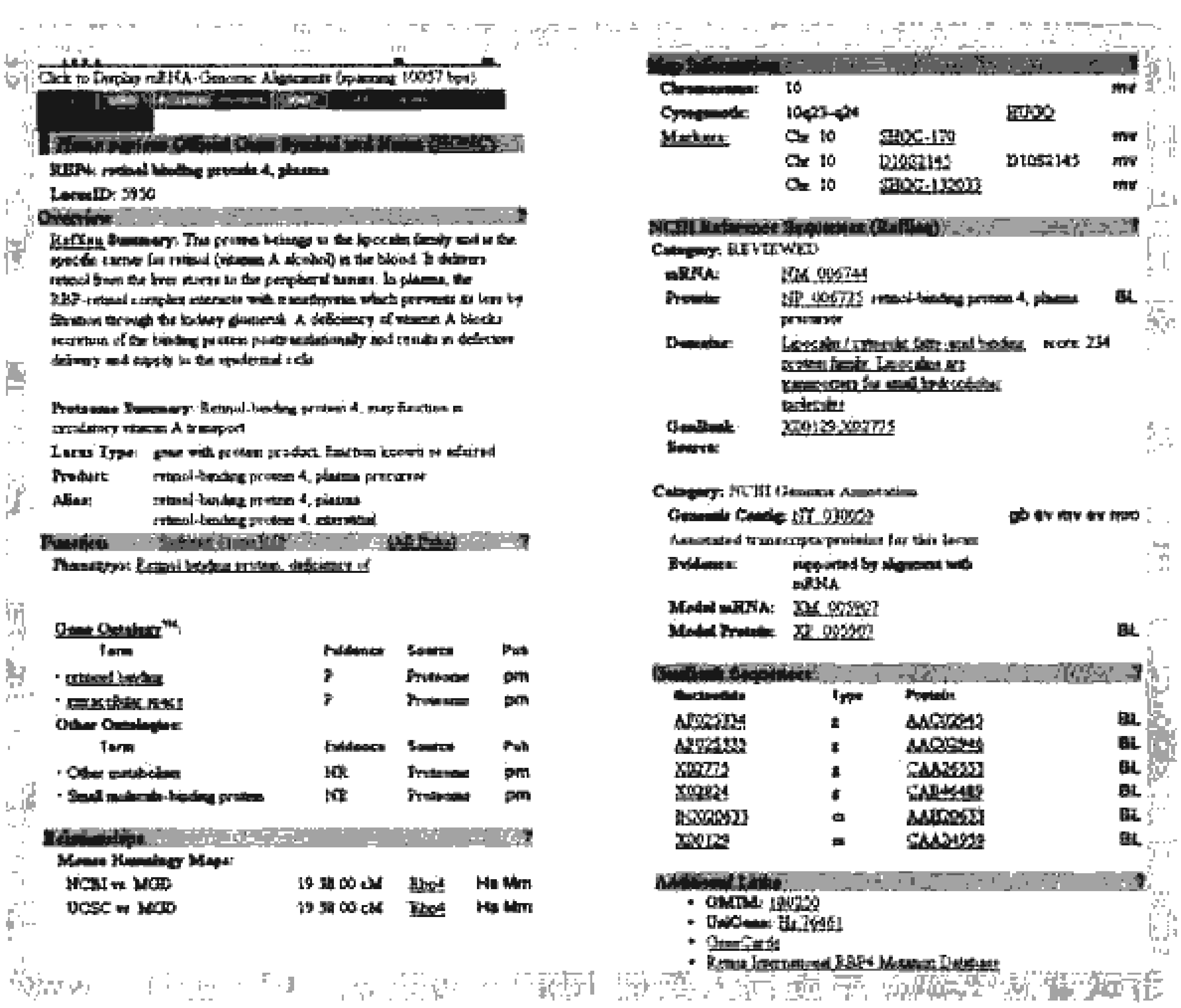


图 2-8　人类 RBP4 的 LocusLink 数据记录。页面的顶部是到各种数据库中的 RBP4 的链接，包括 PubMed、OMIM、UniGene、Map、变体数据库（dbSNP）、HomoloGene（小鼠和大鼠 RBP4 的信息）、基因组数据库 GDB、基因 ontology 数据库和一个 Ensembl 浏览器。我们将在后面的章节中讨论这些数据库。LocusLink 还提供了有关 RBP4 的其他关键的信息，包括被选作代表的 RBP4 的核酸或蛋白的记录

结果还给出GenBank中RBP4的索引号码（核酸和蛋白质），包括X00129和X02775。

2.3.3.2 NCBI参考序列：最好的代表序列

等位基因的变体（如基因中单个碱基的变异）并没有被赋予不同的RefSeq的索引号码。但是OMIM和dbSNP数据库都记录了等位基因变体（第18章）。

参考序列（RefSeq）给出了一个对应于基因或蛋白质的索引号码，对应于最稳定、最被人承认的序列（Pruitt等，2000；Maglott等，2000）。GenBank中对应于某个基因的索引号可能有上百个，但是对应于一个基因的RefSeq记录只有一个，或者在存在可变剪切的情况下一个基因对应不止一个RefSeq记录。这个记录包含了不同层次（预测的、临时性的，或者是总结性的），但是每一个层次中RefSeq记录的目标都是为了使序列记录统一化。

你可以从索引号码来识别RefSeq的记录，如NP_006735（RBP蛋白质）或者NM_006744（RBP的mRNA）。表2-7给出了RefSeq的索引号码格式。

表2-7 RefSeq记录的索引号码格式

分子类型	索引号码	基因组
完整基因组	NC_######	古细菌、细菌、细胞器、病毒
完整的染色体	NC_######	真核生物
完整的序列	NC_######	质粒
基因组contig	NT_######	人（*Homo sapiens*）
基因组草图	NW_######	大鼠Rat
mRNA	NM_######	有限的几个脊椎动物：人、小鼠、大鼠
蛋白质	NP_######	上述所有的基因组

2.3.3.3 UniGene

对质粒的描述见图6-5。

如上文所述，UniGene为每一个基因给出了一组序列。例如，对于RBP4来说，对应于UniGene的索引号码为Hs.418083。这个UniGene条目存储了GenBank中对应于RBP4的所有的记录列表，包括EST。UniGene还给出了图谱信息、同源基因和表达信息（有基因表达的组织列表，也就是在这些组织中获取了cDNA库，用来产生对应于RBP4的EST）。

UniGene和LocusLink的比较。UniGene和LocusLink有一些共同的特点，如与OMIM、同源基因，以及图谱信息的链接。它们都给定了RefSeq的索引号码。UniGene和LocusLink的四个主要区别如下。

LocusLink目前收录了约41000个人类的RefSeq条目（到2003年5月）。其中约有14000个条目属于"编码蛋白质的基因，功能已知或者可推导"。

① UniGene有具体的表达信息：它提供了作为EST记录来源的cDNA库的组织区域分布列表。

② UniGene给出了对应于基因的EST列表，允许使用者对它们进行深入的研究。

③ LocusLink可能提供对一个基因更稳定的描述：如上文所述，UniGene记录可能随着基因组测序的进展而被从数据库中删除。

④ LocusLink比UniGene的记录数目少，但是这些记录却更精确。

2.3.3.4 Entrez

Entrez是一个很流行的用于提取序列信息的工具。类似于SRS（见下文），Entrez是一个查询、提取和显示系统。请试着从Entrez中查询人类的RBP4。可以从NCBI主页（www.

ncbi. nlm. nih. gov）点击 Entrez 或者直接从 Entrez 的地址进入（http://www. ncbi. nlm. nih. gov/Entrez/）。设定你要搜索一个蛋白质，然后输入“retinol-binding protein”，你将得到超过 200 个返回结果（“hits”），这对我们来说显然太多了。如果我们输入“human retinol-binding protein”，可以得到大约 75 个结果。用“retinol-binding protein human interstitial”搜索可以得到 9 个结果，其中一个结果中包含一个 RefSeq 索引号（NP _ 006735）。图 2-9 给出了这条 Entrez 条目的默认形式。FASTA 是另外一种记录蛋白质（或 DNA）的通用格式，如图 2-10 所示。通过改变 Entrez 的显示选项可以很容易地获取各种格式的数据。

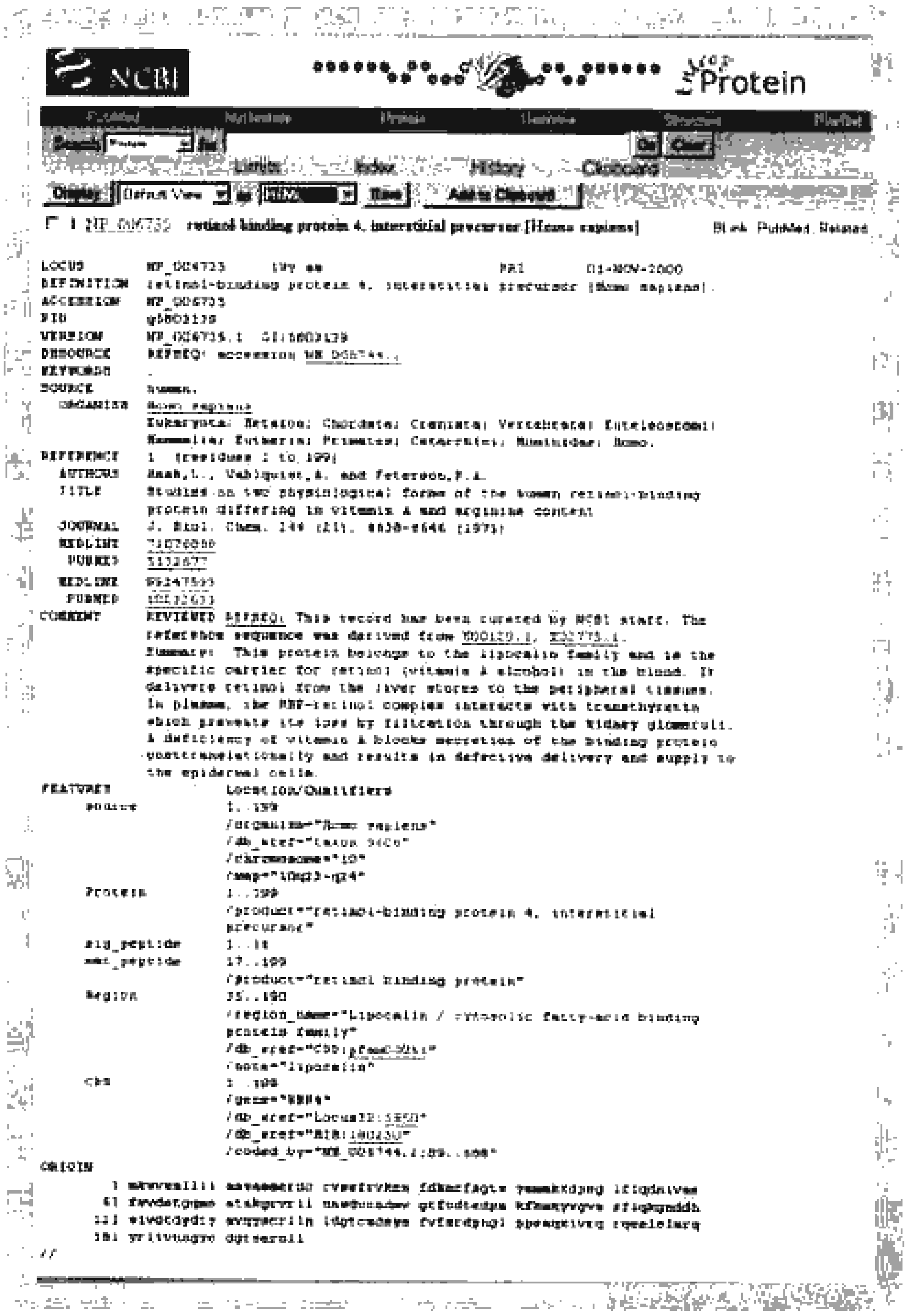


图 2-9　Entrez 给出的一个人类 RBP4 的 GenBank 的蛋白数据条目。这是蛋白质的典型的条目（DNA 记录也是很相似的）。关键的信息还包括蛋白的长度（199 个氨基酸序列）、分类信息（PRI 或者 primate）、索引号码（NP _ 006735）、物种信息（*H. sapiens*）、参考文献、RBP 功能的注释，以及类似编码序列（CDS）的信息。图的下方给出了蛋白的氨基酸序列。在页面上部的 Display option 选项框可以用来改变输出结果的格式，如 FASTA（图 2-10）

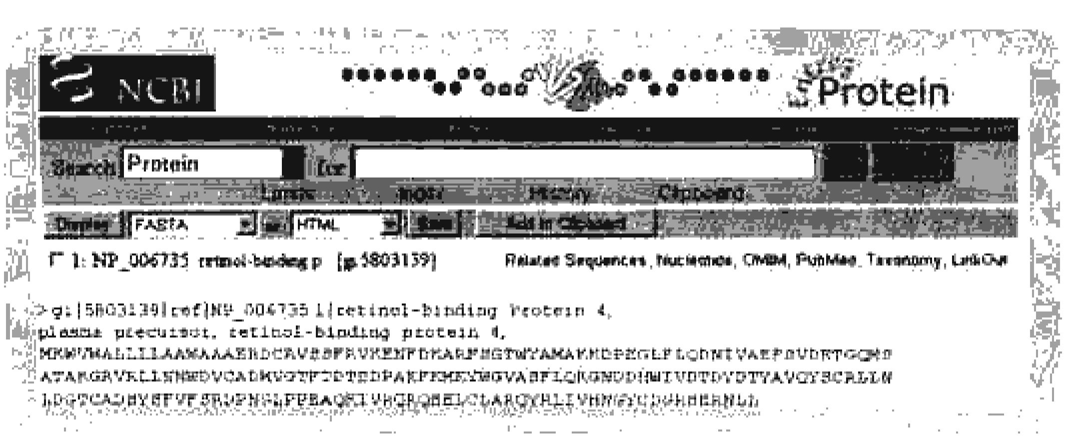


图 2-10　人类 RBP4 的蛋白质数据条目可以用 FASTA 格式显示。这可以通过改变 Entrez 蛋白质记录的“Display”下拉菜单来改变输出的数据格式。许多软件都使用 FASTA 格式，我们将在后面的章节中讨论 FASTA 格式的数据

注意到你得到的搜索结果会随着数据库的更新而发生变化。

FASTA是一个用于序列比对的程序（在第3章中介绍），同时FASTA也表示一种序列的格式（我们将在第4章中做进一步介绍）。

NCBI的网站还提供了一个非常有价值的Entrez使用教程，其地址是http://www.ncbi.nlm.nih.gov/Education/index.html。Entrez的许多特性中，非常重要的一个就是用“limits”可以把搜索结果限制在你真正感兴趣的范围内。以搜索“RBP”为例，搜索Entrez的核酸数据将返回550个条目。选择“limits”，然后改变下拉菜单的选项到“Only from RefSeq”，然后选择“go”，你将得到65个条目。这样你所要搜索的信息就更容易被找到了。在这个例子中，属于*H. sapiens*的RBP4的DNA序列可以轻易地从列表中被找出来（RefSeq索引号码为NM _ 006744）。

2.3.3.5 欧洲生物信息学研究所（EBI）和Ensembl

EBI通过EMBL核酸数据库提供了序列搜索的服务（Hingamp等，1999）。通过它的序列提取系统——SRS6（Etzold等，1996），我们可以用十几种不同的方法（如用关键字）搜索我们想要的序列。NCBI提供了搜索GenBank的工具Entrez，而EBI的SRS6也实现了同样的数据库搜索功能。EBI还资助了Ensembl项目，Ensembl是一个用于对人类基因组进行生物信息学分析的非常完备的网站。用“rbp4”搜索Ensembl站点可以得到一个指向RBP4蛋白和*RBP4*基因的链接；我们将在后面的章节中继续讨论Ensembl资源。刚刚取得的这个条目含有大量有关RBP4的属性的信息，包括序列名称、DNA序列以及与其他许多数据库资源的交互链接（包括OMIM和RefSeq）。

Ensembl（http://www.ensembl.org）是由EBI和Wellcome Trust Sanger Institute（http://www.sanger.ac.uk/）合作进行维护的。EMBL可以通过http://www.ebi.ac.uk/embl/进行访问。

2.3.3.6 ExPASy

用于获取蛋白质序列和相关数据的最有用的资源之一就是ExPASy，也就是蛋白质专家分析系统。查询页面有四个长方形的输入框（图2-11），每一个都对应一个下拉菜单，默认选项都是“AllText”。在第一个输入框中，输入“retinol-binding”（注意必须保证查询用的关键字是一个单词）。在第二个输入框中输入“human”，然后把对应的输入菜单项改为“organism”，然后点击“do query”，你将得到10个条目。点击我们真正感兴趣的那个链接（SWISS _ PROT：RETB _ HUMAN P02753）。

有一条输出结果包含一条SwissProt记录。这提供了非常有用的、组织得非常好的信息，包括别名和索引号码，文献链接，功能数据和有关细胞定位的信息，链接到RBP蛋白和基因的GenBank条目和其他数据记录的链接，以及与许多别的数据库的链接（如OMIM、InterPro、Pfam、Prints、GeneCards、PROSITE和蛋白质二维凝胶电泳数据库）。我们将在后面的章节讨论这些数据库（第8～第10章）。数据的记录中还包括了RBP的属性，通过点击其中任何一个属性链接，你将看到一条用颜色把你点击的属性显示出来的蛋白质序列。

ExPASy是瑞士生物信息研究所的蛋白质组服务器（http://www.expasy.ch/），也是另外一个序列查询系统（SRS）的入口。读者可以访问http://www.expasy.ch/srs5/，点击“Start a new SRS section”，然后点击“continue”。SRS是由Lion Biosciences公司开发的（http://www.lionbioscience.com），还有一系列公共的SRS服务器可以从http://downloads.lionbio.co.uk/publicsrs.html中取得。

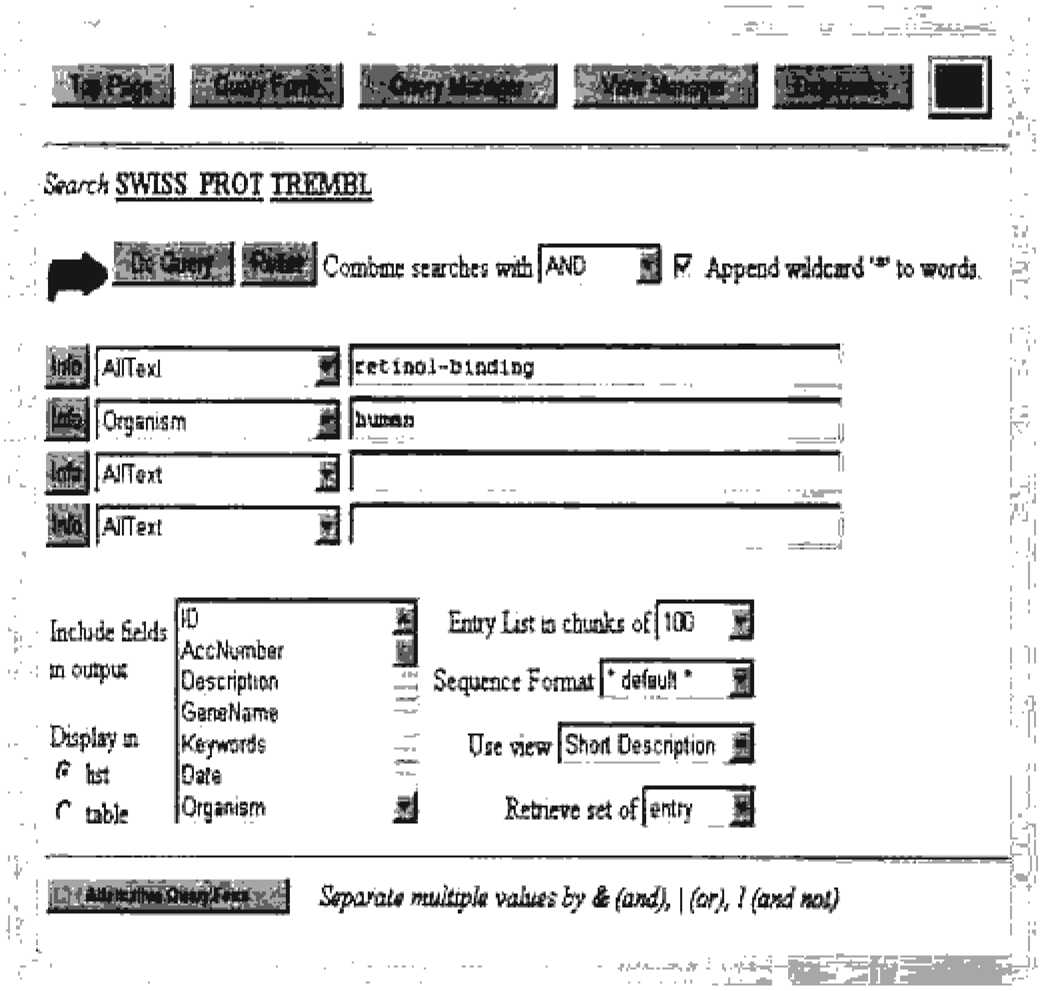


图 2-11 蛋白质分析专家系统（ExPASy）的序列查询系统（SRS）的界面（http://www.expasy.ch/srs5/）。这个网站提供了蛋白质分析的一个重要资源。你可以通过欧洲生物信息研究所的网站访问 SRS 系统（http://srs6.ebi.ac.uk）

2.3.4 如何进入序列数据的例子：HIV *pol*

考虑到逆转录酶，HIV-1 的 RNA 依赖 DNA 聚合酶（Frankel 和 Young，1998）。编码这个逆转录酶的基因叫做 *pol*（polymerase）。怎样得到它的 DNA 和蛋白质序列呢？

到 NCBI 的 Entrez 的蛋白质分项，将搜索限制在基因组数据库（genome database）中，然后输入 "hiv-1"。我们将得到两个结果：HIV-1 和另外一个病毒。HIV-1 的索引号码（NC _ 001802）是一个 RefSeq 号码，对应于完整的病毒基因组。点击这个 HIV-1 的链接，你将看到一个含有九个 HIV-1 开放阅读框（ORF）的图形（图 2-12）。其中之一的 *pol* 就是我们要寻找的聚合酶的基因。接着，我们有好几种方法可以得到它的蛋白质序列。第一种方法：点击 NC _ 001802 这个链接，蛋白质序列就在对应的 Entrez 条目中，其后紧跟着蛋白质的索引号码（NP _ 057849.2）。NP _ 057849.2 是一个 RefSeq 条目，其中的后缀 ".2" 表示这是 GenBank 中这条序列的第二个版本。这个蛋白质含有 1003 个氨基酸。第二种方法：用左边导航条中的 "graphical view" 或者 "coding regions" 来获得 HIV-1 中的蛋白质的图形。逆转录酶的信息同样也可以用这种方法获得。

我们现在介绍取得这条蛋白质序列的另一种方法。到 NCBI 网站上的 Entrez 的蛋白质的子项（http://www.ncbi.nlm.nih.gov/Entrez），将搜索限制在蛋白质数据库中，然后输入 "reverse transcriptase"。我们将得到超过 4 万个结果。为了使搜索结果更符合我们的要求，改用 "reverse transcriptase hiv-1" 进行搜索，这

虽然我们这里提到了几个获取序列的方法，但实际上还存在着许多有用的服务器。例如 Protein Information Resource（PIR）也提供了序列搜索的服务（Wu 等，2002）。PIR 在蛋白质的功能注释方面做了重要的工作。

PIR 是美国国家生物医学研究基金会（http://pir.georgetown.edu）的一个分部。2003 年 2 月发布的 PIR 版本含有约 140 万个条目。PIR 是由 Margaret Dayhoff 创建的，我们将在第 3 章中讨论她的工作。

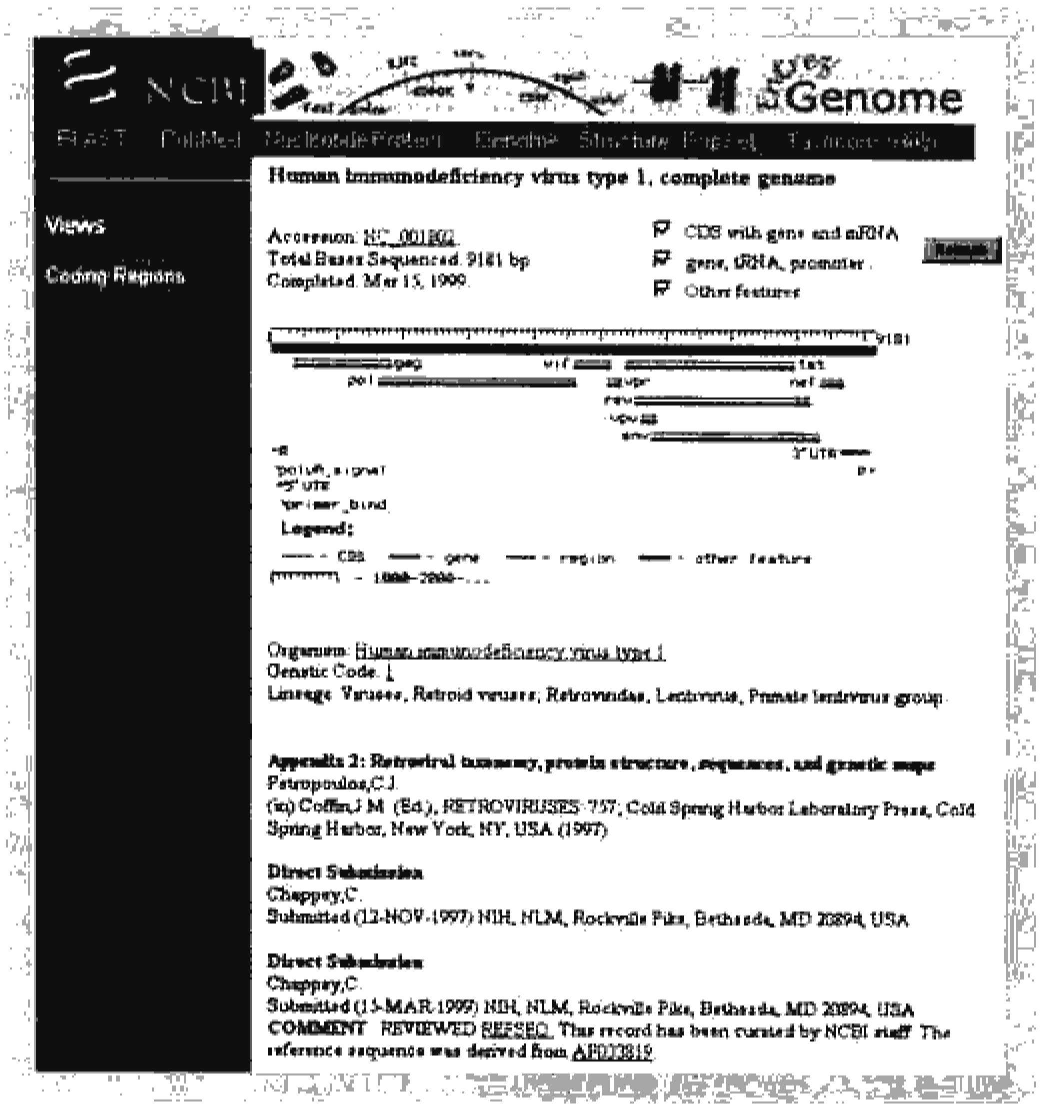


图 2-12 人类获得性免疫缺陷病毒（HIV-1）的 Entrez 基因组记录（源自 http://www.ncbi.nlm.nih.gov/Entrez）。以 NC_开头的 RefSeq 条目是完整基因组的记录（表 2-6）。Entrez 记录中的关键信息包括索引号码、基因组长度（9181bp）、基因组中基因/蛋白质的图形概览、分类信息以及参考文献。左边的一栏中，"Coding Regions" 选项提供了 HIV-1 的 9 个基因（包括 *pol*）的索引号码

联合蛋白质数据库（The United Protein Database，UniProt）是一个新的蛋白质中心资源，它包括了 PIR 提供的数据库、欧洲生物信息学研究所（EBI）和瑞士生物信息学研究所（SIB）的数据库。UniProt 的网站是 http://www.uniprot.org。

你可以通过 http://www.ncbi.nlm.nih.gov/Entrez 进入 Entrez 网站。注意到这些数据库更新得非常快，因此我们在书中提到的搜索结果可能随着时间的推移发生变化。NLM 的网址是 http://www.nlm.nih.gov。

次我们得到 15000 个结果，这显然还是太多了。选择 "limits"，然后把 "only from" 下拉菜单改到 "RefSeq"，然后点击 "Go"。这时候我们得到的结果是可以手工处理的，包括我们想要的 NP_057849 这个索引号码。

注意到其他的 NCBI 数据库不适合用来寻找病毒的逆转录酶序列：UniGene 不收录病毒记录，而 OMIM 只限于人类的数据。但是 UniGene 和 OMIM 确实保存有与 HIV 相关的基因的记录，如真核生物的逆转录酶。

最后，我们还可以通过 SRS 获取 HIV-1 逆转录酶的序列。选择 SwissProt 数据库作为目标数据库源。在四个选择对话框中，把其中一个设为 "organism" 和 "HIV-1"，然后把另一个设置为 "All Text" 和 "reverse"。然后点击 "Do query"，这时候我们将得到 10 多个符合要求的条目，其中很多是序列片段，可以被忽略。得到的第一个条目是 SWISS_PROT：POL_HV1A2（SWISS_PROT 索引号码 P03369），这是一个具有 1003 个氨基酸的蛋白质。点击 SwissProt 的链接，然后找到其中与 "NiceProt" 的链接。这些信息包含着记录的条目和修改日期、蛋白质的名称和别名、参考信息（PubMed 链接）、注释（包括一个简要的功能描述）、与许多其他很有用的数据库的交叉链接、关键词列表、一些类似预测的二级结构的特征信息，以及用单个字母表示氨基酸的蛋白质序列编码和预测的蛋白质分子质量。

我们也可以通过欧洲生物信息学研究所的 SRS (http://srs.ebi.ac.uk/) 来进入 MEDLINE。在 NCBI 网站上提供了一个 PubMed 教程，网址是 http://www.nlm.nih.gov/bsd/pubmed_tutorial/m1001.html。

2.3.5 获取生物医学文献

美国 NLM（National Library of Medicine）是全球最大的医学图书馆。1971 年，NLM 开发了 MEDLINE（medical literature，analysis，and retrieval system online，在线生物医学文献分析和提取系统）——一个文献数据库。MEDLINE 包含了 1100 万个在生命科学期刊中发表的参考文献记录，来自 70 多个国家的 4300 个生物医学期刊的引用记录。NCBI 开发的 PubMed 的网站（http://www.ncbi.nlm.nih.gov/PubMed/）提供了免费的 MEDLINE 查询服务。虽然 MEDLINE 和 PubMed 都提供了参考文献引用的信息，但 PubMed 还给出了在线全文期刊文章的链接。PubMed 还提供了 NCBI 中生物数据库的集成链接。这些数据库收录了 DNA 和蛋白质序列、基因组图谱数据和蛋白质三维结构信息。

有关 MEDLINE 的增长的描述也可以从网站上获得（http://www.nlm.nih.gov/bsd/medline_growth.html）。虽然 MEDLINE 记录的文献由许多不同的语言的文章组成，但 2000 年其中记录的英文文章所占的比例为 88%，而在 1966 年这个数字是 59%（http://www.nlm.nih.gov/bsd/medline_lang_distr.html）。

2.3.5.1 PubMed 集成化和文献访问免费化

生物医学界正展开一项关于获取文献信息方式的讨论。目前，许多期刊都提供了免费或者通过互联网进行订购后访问电子版文献的服务。类似研究文献图书馆联盟（Association of Research Libraries，ARL）的一些组织正密切关注着文献出版向电子化过渡的趋势。目前已经有几千个期刊提供了在线服务。越来越多的在线文章都附带了类似于分子数据（如全基因组序列或基因表达谱数据），或者用来说明文章的录像剪辑这样的附属材料。PubMed 集成化提供了一个生物学文献的中心存储站（Robert，2001）。所有这些文献都是经过同行评审，并且在相应的期刊中同步出版的。

美国国家医学图书馆通过 NLM 网站也提供了到 PubMed 的入口（http://gateway.nlm.nih.gov）。这个综合网站还提供了从 PubMed 无法访问到的一些 NLM 下属的数据库的链接，如到会议论文摘要以及一个医学百科全书的链接。

2.3.5.2 PubMed 搜索的例子：RBP

用 RBP 作关键字在 PubMed 中搜索可以得到约 1200 条记录。图 2-13 说明了 PubMed 中的逻辑运算符的基本用法。有许多附加的方法可以用来对搜索进行限制。点击“limits”，然后试着把输出文献限制在那些能通过 PubMed 中心免费取得全文的文献。

ARL 网址是 http://www.arl.org/scomm/edir/index.html。

Medical Subject Headings（MeSH）提供了一个方便的聚焦或扩展搜索的方法。MeSH 是一个经过特别整理的词汇字典，它收录了 22000 个条目。在 PubMed 网站中，在左边的一列中点击“MeSH Database”，然后输入“retinol-binding protein”。搜索的结果提示了一系列可能与 RBP 有关的研究领域。通过增加 MeSH 限制条目，我们可以把搜索限制在你所需要的特定的信息中。Lewitter（1998）与 Fielding 和 Powell（2002）讨论了有效的 MEDLINE 搜索的方法，比如避免在 MeSH 条目的不一致性以及在灵敏性（找到相关文献）和准确性（去除不相关的引用）中寻找平衡点的方法。例如，对于一个还没有很好的被索引的研究领域来说，把一个文字关键词和一个 MeSH 条目混合起来使用是很好的搜

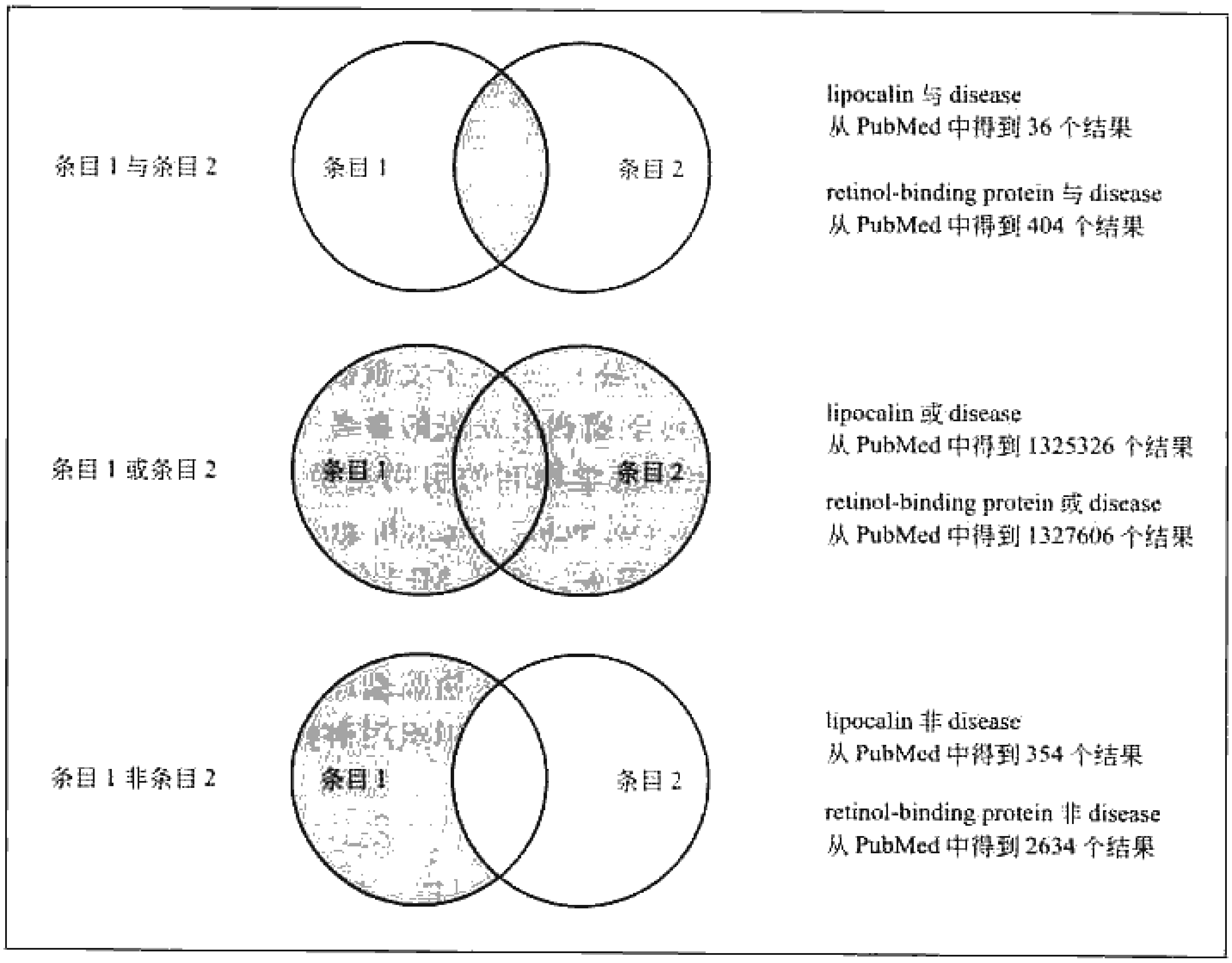


图 2-13 用维因图（Venn Diagrams）表示两个假设的搜索条目（条目1和条目2）的逻辑操作符 AND、OR 和 NOT。AND 表示要搜索的条目1和条目2在每个结果中都出现，OR 允许最后的结果包含至少其中的一个条目，而 NOT 表示排除包含后面一个搜索条目的结果。图中灰色的区域表示返回的结果，例子是用“lipocalin”或“retino-binding protein”在 PubMed 中搜索的结果，逻辑操作符间接影响返回的结果

NLM 提供的 MeSH 网站地址是 http://www.nlm.nih.gov/mesh/meshhome.html。

索办法。使用通配符也是一个很好的办法，如搜索“therap*”时引入一个通配符将搜索出类似 therapy、therapist 和 therapeutic 这样的条目。

2.4 展望

生物信息学是一门非常年轻的学科，生物信息学兴起的标志是数据库中生物信息的大量积累。三个主要的 DNA 数据库——GenBank、EMBL 和 DDBJ 每年都新增加几百条新的序列，对应约数十亿的核苷酸。

在本章中，我们讨论了获取 RBP4 和 HIV *pol* 基因的 DNA 和/或蛋白质序列。除了三个主要的数据库之外，网上还有许多有用的资源。正确的搜索方法不止一个——许多搜索途径都是可行的。我们进一步看到本章中讨论的这些资源——NCBI、ExPASy、EBI/EMBL 和 Ensembl 都是相互链接的，数据库之间都提供了指向对方的链接。

2.5 常见问题

在获取序列和文献信息的时候常伴随着许多常见的问题。在任何一次搜索之前，最重要

的第一步是确定你的目标：例如，确定你要搜索的是蛋白质还是DNA序列。在进行数据库搜索的时候经常碰到的一个难题是得到了太多的返回结果，它可通过学习如何用合适的限制来确定搜索范围来解决。

2.6 网络资源

表2-8和表2-9提供了有关核酸和蛋白质，以及相关文献的一些网络资源。

表2-8 核酸和蛋白质信息数据库

数据库名称	网址
GenBank	http://www.ncbi.nlm.nih.gov/Genbank
DNA Database of Japan(DDBJ)	http://www.ddbj.nig.ac.jp/
EMBL/EBI(European Molecular Biology Lab/European Bioinformatics Institute)	http://www.ebi.ac.uk/
Ensembl	http://www.ensembl.org/
Protein Information Resource	http://pit.georgetown.edu/
SRS at ExPASy(ExPASy提供的SRS服务)	http://www.expasy.ch/srs5
SRS at DDBJ(DDBJ提供的SRS服务)	http://srs.ddbj.nig.ac.jp/index-e.html

表2-9 文献数据库

数据库名称	网址
PubMed	http://www.ncbi.nlm.nih.gov/PubMed/
Public Library of Science(公共科学图书馆)	http://www.publiclibraryofscience.org/

2.7 讨论题

你可以通过本书所在的网站(http://www.bioinfbook.org)获取按章节组织起来的所有这些数据库的网站。

1. 数据库中会出现哪些种类的错误？这些错误如何来评估？

2. 假设有几千个独立的研究人员向GenBank中提交数据，在GenBank中如何进行数据质量调控？

2.8 习题

1. 人类的蛋白质中有哪些超过了300000Da？(提示：使用Entrez)

2. 假定你对乳腺癌相关的基因感兴趣，请问这些基因有哪些？其中的DNA和蛋白质的索引号码是什么？尝试下面5种方法分别搜索你要的信息：PubMed、Entrez、LocusLink、OMIM和ExPASy的SRS。

3. 如你想更进一步了解*BRCA1*这个基因——与乳腺癌相关的肿瘤抑制因子，请说明这个基因的序列来自于哪些物种？

4. ATP（腺苷三磷酸）结合区（ABC）是许多所谓的ABC转运蛋白质中通用的一个蛋白质结构域的例子。你知道人类中带有ABC这个结构域的蛋白质大概有多少个？细菌中这样的蛋白质又有多少个呢？在讨论题2中的那5个数据库中哪个对于你得到一个确定的ABC结构域的定义更有用呢？(我们将在第8章中讨论更多的资源)

5. 找一个lipocalin蛋白的索引号码（如retinol-binding protein，lactoglobulin，任何细菌的lipocalin，glycodelin或者odorant-binding protein）时，首先用LocusLink，然后用

UniGene，接着用 OMIM 和 Entrez 核酸数据库，最后用 Entrez 蛋白质序列库，哪个途径更有效？这个蛋白质有什么功能？

6. 基于文字的分子信息搜索的三个主要的工具是：美国国家生物技术中心（NCBI）的 PubMed，Entrez 和 OMIM 工具（http://www.ncbi.nlm.nih.gov）；欧洲生物信息学研究所（EBI）的序列提取系统（SRS）（http://srs.ebi.ac.uk）或者与它相关的 SRS 站点（http://www.expasy.ch/srs5）；京都大学和东京大学的 GenomeNet 工具 DBGET（http://www.genome.ad.jp/dbget/dbget2.html）及文献数据库 LitDB。当你准备研究西尼罗河病毒（West Nile Virus）时，用这些资源分别进行搜索时得到了怎样的结果？

7. 如想知道在 Bioinformatics 这个杂志中有哪些和病毒相关的论文，用 PubMed 搜索你要的论文。

2.9 自测题

1. 以下哪一个是 RefSeq 中 mRNA 条目的索引号码？

(a) J01536；

(b) NM_15392；

(c) NP_52280；

(d) AAB134506

2. 目前 UniGene 中有多少人类的簇？

(a) 约 8000；

(b) 约 40000；

(c) 约 100000；

(d) 约 150000

3. 你想研究一个基因，你想确定它在哪些组织中表达，下面哪个资源可以帮你最直接地获取这类信息？

(a) UniGene；

(b) Entrez；

(c) LocusLink；

(d) PCR

4. 下面哪个物种没有收录在 LocusLink 中？

(a) Mouse；

(b) Fly；

(c) Human；

(d) *Escherichia coli*

5. 一个基因可能对应两个 UniGene 簇吗？

(a) 是；

(b) 否

6. 下面那种数据库来源于 mRNA 信息？

(a) dbEST；

(b) PDB；

(c) OMIM；

(d) HTGS

7. 下面哪个数据库可以用来获取有关人类疾病的文字信息？

(a) EST；

(b) PDB；

(c) OMIM；

(d) HTGS

8. RefSeq 和 GenBank 有什么区别？

(a) RefSeq 包括了由许多独立的实验室和测序项目提交的公开的 DNA 序列；

(b) GenBank 提供了非冗余序列；

(c) GenBank 序列来源于 RefSeq；

(d) RefSeq 来源于 GenBank，它提供了非冗余的序列信息

9. 如果你需要文献信息，下面哪个数据库是你要访问的最佳站点？

(a) OMIM；

(b) Entrez；

(c) PubMed；

(d) PROSITE

10. 比较从 Entrez 和 ExPASy 中提取有关蛋白质序列信息的方法，下列哪种（或哪

几种）说法是正确的？

（a）因为 GenBank 的数据比 EMBL 更多，Entrez 给出的搜索结果更完整；

（b）搜索结果很可能是一样的，因为 GenBank 和 EMBL 的序列数据实际上是一样的；

（c）搜索结果应该是可以比较的，但是 ExPASy 中的 SwissProt 记录将用一个不同的输出格式给出不一样的信息

2.10 推荐读物

生物信息学数据库正在非常迅速地发展起来。每年 1 月份，"Nuclei Acids Research" 杂志的第 1 期都会发表约 100 篇简要介绍各个数据库的文章，其中就包括 NCBI（Wheeler 等，2002）、GenBank（Benson 等，2002）和 EMBL（Stoesser 等，2002）。特别需要提到的是，Andreas Baxevanis（2002）发表的分子生物数据库集合（molecular biology database collection）。覆盖面非常广。随着与 DNA、RNA 和蛋白相关的序列数据的不断测定，许多专业的数据库被创造出来，用来分析分子、细胞、组织和器官的结构和功能。分子生物学数据库集合目前罗列了近 300 个数据库（http://nar.oupjournals.org/cgi/content/full/31/1/1/DC1）。

参 考 文 献

Adams, M. D., et al. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252**, 1651-1656 (1991)

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).

Altschul, S. F., et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).

Baxevanis, A. D. The Molecular Biology Database Collection: 2002 update. *Nucleic Acids Res.* **30**, 1-12 (2002).

Benson, D. A., et al. GenBank. *Nucleic Acids Res.* **30**, 17-20 (2002).

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365-370 (2003).

Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. dbEST-database for "expressed sequence tags." *Nat. Genet.* **4**, 332-333 (1993).

Etzold, T., Ulyanov, A., and Argos, P. SRS: Information retrieval system for molecular biology data banks. *Methods Enzymol.* **266**, 114-128 (1996).

Fielding, A. M., and Powell, A. Using Medline to achieve an evidence-based approach to diagnostic clinical biochemistry. *Ann. Clin. Biochem.* **39**, 345-350 (2002).

Fleischmann, R. D., et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512 (1995).

Frankel, A. D., and Young, J. A. HIV-1: Fifteen proteins and an RNA. *Annu. Rev. Biochem.* **67**, 1-25 (1998).

Hamosh, A., et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **30**, 52-55 (2002).

Hingamp, P., van den Broek, A. E., Stoesser, G., and Baker, W. The EMBL Nucleotide Sequence Database. Contributing and accessing data. *Mol. Biotechnol.* **12**, 255-267 (1999).

Lewitter, F. Text-based database searching. *Bioinformatics: A Trends Guide* **1998**, 3-5 (1998).

Maglott, D. R., Katz, K. S., Sicotte, H., and Pruitt, K. D. NCBI's LocusLink and RefSeq. *Nucleic Acids Res.*, **28**, 126-128 (2000).

O'Donovan, C., et al. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinform.* **3**, 275-284 (2002).

Olson, M., Hood, L., Cantor, C., and Botstein, D. A common language for physical mapping of the human genome. *Science* **245**, 1434-1435 (1989).

Pruitt, K. D., Katz, K. S., Sicotte, H., and Maglott, D. R. Introducing RefSeq and LocusLink: Curated human genome resources at the NCBI. *Trends Genet.* **16**, 44-47 (2000).

Roberts, R. J. PubMed Central: The GenBank of the published literature. *Proc. Natl. Acad. Sci. USA* **98**, 381-382 (2001).

Sarton, G. *Appreciation of ancient and medieval science during the renaissance* (1450-1600). A. S. Barnes & Co., New York, 1955.

Singer, C. *The Fasciculo di Medicina Venice 1493*. R. Lier and Co., Florence, 1925.

Stoesser, G., et al. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* **30**, 21-26 (2002).

Westbrook, J., et al. The Protein Data Bank: Unifying the archive. *Nucleic Acids Res.* **30**, 245-248 (2002).

Wheeler, D. L., et al. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.* **30**, 13-16 (2002).

Wu, C. H., et al. The Protein Information Resource: An integrated public resource of functional annotation of proteins. *Nucleic Acids Res.* **30**, 35-37 (2002).

Adrenocorticotropin (ACTH)

The complete amino acid sequences are known for corticotropins isolated from the anterior pituitary glands of three different species, pig, beef, and sheep. The structure of sheep ACTH was discussed in the last chapter, and the sequences shown in Table 9 include only those areas of the three molecules where differences are to be found. Although some difference between the content of amide nitrogen groups has been reported for the three species, these are not included in the figure since it has not been possible to rule out, with certainty, the possibility that these variations are due, in part, to the rigors of the isolation and purification techniques employed.

TABLE 9
Variations in Amino Acid Sequences Among Different Preparations of ACTH

Preparation	Species	Residue No. 25 26 27 28 29 30	31 32 33
β-Corticotropin	sheep, beef*	Ala.Gly.Glu.Asp.Asp.Glu	Ala.Ser.Glu.NH_2
Corticotropin A	pig	Asp.Gly.Ala.Glu.Asp.Glu	Leu.Ala.Glu

* Identity with sheep hormone not absolutely certain but very probable as judged from the nearly complete sequence analysis by J. S. Dixon and C. H. Li (personal communication to the author).

Two points are of particular interest in regard to the sequences shown. First, the corticotropins of sheep and beef are identical and differ from that of the pig. This finding is consonant with the closer phylogenetic relationship of sheep and cows to each other than of either to pigs. Second, chemical differences are found only in that portion of the ACTH molecule which has been shown to be unessential for hormonal activity. Genetic mutations leading to such differences might, therefore, not be expected to impose significant disadvantages in terms of survival, and these genes could become established in the gene pools of the species.

Melanotropin (MSH)

Melanotropin, like the other hormones considered in this chapter, is a typically chordate polypeptide. Indeed, the demonstration of melanocyte-stimulating activity in extracts of tunicates constitutes an

双序列比对指匹配两条蛋白质或DNA序列。第一个测序的蛋白包括胰岛素（Frederick Sanger及其同事，见图11-1）和球蛋白。上图来自诺贝尔奖获得者Christian Affinsen所著的《进化的分子基础》（1959年，153页）。它显示了羊或牛（上部）和猪（下部）的促肾上腺皮质激素（TACH）的一部分的双序列比对结果。这类由手工得到的比对结果使我们认识到蛋白质氨基酸序列的关系可以反映不同物种在系统发育上的关系。进一步，双序列比对结果可以揭示蛋白质的哪一部分对其生物学功能可能是重要的。授权使用

第3章

双序列比对

3.1 引言

有关一个基因或蛋白质的一个基本问题是它是否和其他的基因或蛋白质相关。两个蛋白质在序列水平上的相关性预示着它们的同源性。相关性也暗示它们可能会有相同的功能。通过分析许多DNA和蛋白质序列，我们有可能确定一群分子之间共享的结构域（domain）或模体（motif）。这些基因或蛋白质的相关性分析是由序列比对来完成的。当我们完成对许多基因组的测序以后，寻找物种内和物种间蛋白质的相关性对于我们理解生命来说变得越来越重要。

如果两条基因（或者蛋白质）由共同的祖先进化而来，它们是同源的。

在这一章我们将介绍双序列比对。我们将从进化的角度来描述两条序列之间的氨基酸（或者核苷酸）是如何被比对和相互比较的。然后我们会描述双序列比对的算法和程序。

3.1.1 蛋白质比对：通常比DNA比对具有更丰富的信息

如果要在DNA序列比对和对应编码的蛋白质序列比对之间选择，通常是比较蛋白质序列更有用。其理由有：DNA序列的许多改变（特别是在密码子的第三位）不会改变对应的氨基酸；而且，许多氨基酸有相似的生物物理性质（如赖氨酸和精氨酸都是碱性氨基酸）。在比对时可以用一打分系统（见本章）来描述这些相关（但不匹配）的氨基酸之间的重要相关性。从这个方面来考虑，DNA序列提供的信息较少些。蛋白质序列比对能确定来自各物种的同源序列，只要这些物种有一个10亿年前（billion years ago，BYA）的共同祖先（例如谷胱甘肽转移酶）（Pearson，1996）。对比之下，DNA序列比对一般情况下只能回溯到最多6亿年前（600million years ago，600MYA）。

当要分析一个核苷酸编码序列时，研究它对应的蛋白质序列更可取。在第4章（BLAST搜索），我们将发现能很容易地将DNA转换成蛋白质。例如，tblastn（可以从NCBI BLAST网站得到）允许从一DNA数据库中搜索和给定与蛋白质序列相关的DNA序列（见第4章）。这种查询是通过把每条DNA序列翻译成它可能编码的六条蛋白质序列来实现的。

然而，在一些情况下，比较核苷酸序列比较合适。例如，在确认给定DNA序列和

DNA 数据库中的序列的一致性时，在搜索多态性时，在分析所克隆的 cDNA 片段的一致性时，或者在其他许多情况下，核苷酸的比较就显得重要。

3.1.2 同源性、相似性、一致性的概念

让我们来看一下 lipocalin 蛋白质家族。我们从人 RBP4 [accession number（访问序号）NP _006735] 和牛 β-乳球蛋白（accession number P02754）开始（它们相关性较远但有统计显著性）。RBP4 的 accession number 来自于 LocusLink，β-乳球蛋白的 accession number 来自于 Entrez，因为 LocusLink 中现在还没有这条序列。

如果两条序列有一个共同的进化祖先，那么它们是同源的。这里不存在同源性（homology）的程度问题，两条序列之间要么是同源的，要么是不同源的（Reeck 等，1987；Tautz，1998）。同源蛋白总是在三维结构上有显著的相似性（similarity）。RBP4 和 β-乳球蛋白有非常相似的 X 射线结晶学测得的结构（图 3-1）。当两条序列同源时，它们的氨基酸或核苷酸序里通常有显著的一致性（identity）。因此，同源性是序列同源或不同源的一种论断，而一致性和相似性是一种描述序列相关性的量。值得一提的是，两蛋白质（核苷酸）即使没有统计上显著的一致性，它们也可能是同源的。例如，在 lipocalin 家族，所有成员都是同源的，但是有些序列变化较大，以至于它们没有可检测到的序列一致性。通常，两蛋白质之间三维结构的趋异比氨基酸序列一致性的趋异要慢（Chothia 和 Lesk，1986）。识别此类同源性是一个特别有挑战性的生物信息学问题。

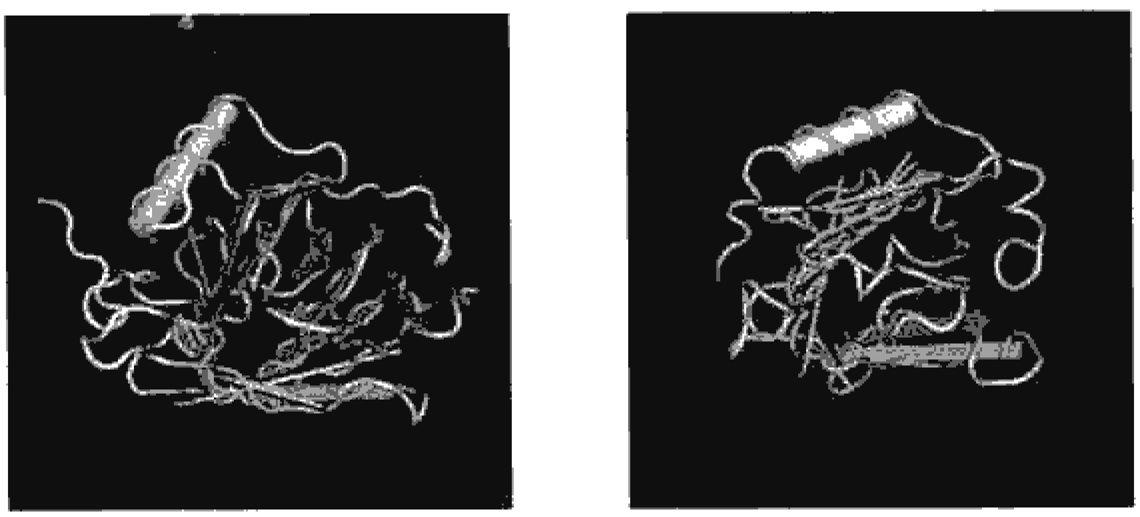


图 3-1 两个 lipocalin 家族成员的三维结构：人 RBP4（左图）和牛 β-乳球蛋白（右图）。图像由程序 Cn3D（见第 9 章）所绘。两个蛋白质具同源性（有一共同祖先），它们有非常相似的三维结构，包括一配体的结合口袋和八个反平行的 β 链组成的 β 片层。但是，两个蛋白质氨基酸序列的双序列比对表明它们只有非常有限的氨基酸一致程度。accession number 为：MMDB Id：934，PDB Id：IFEM（RBP）；MMDB Id：11969，PDB Id：1BSQ（β-乳球蛋白）

同源的蛋白质其同源性特征可能是直系同源（orthologous）或者是旁系同源（paralogous）。直系同源序列是不同物种内的同源序列，它们来自于物种形成时的共同祖先基因。图 3-2 显示了由 RBP 直系同源序列组成的一棵树，其中有一人 RBP 基因和大鼠基因。人类和啮齿类大约在 75MYA 分化，那时单个祖先 RBP 基因随物种形成开始分化。直系同源基因被认为有相似的生物学功能。这个例子中，人和鼠的 RBP 都负责转运血清中的维生素 A。旁系同源基因是通过类似基因复制的机制产生的同源序列。例如，人血浆 RBP 蛋白和另一个载运蛋白

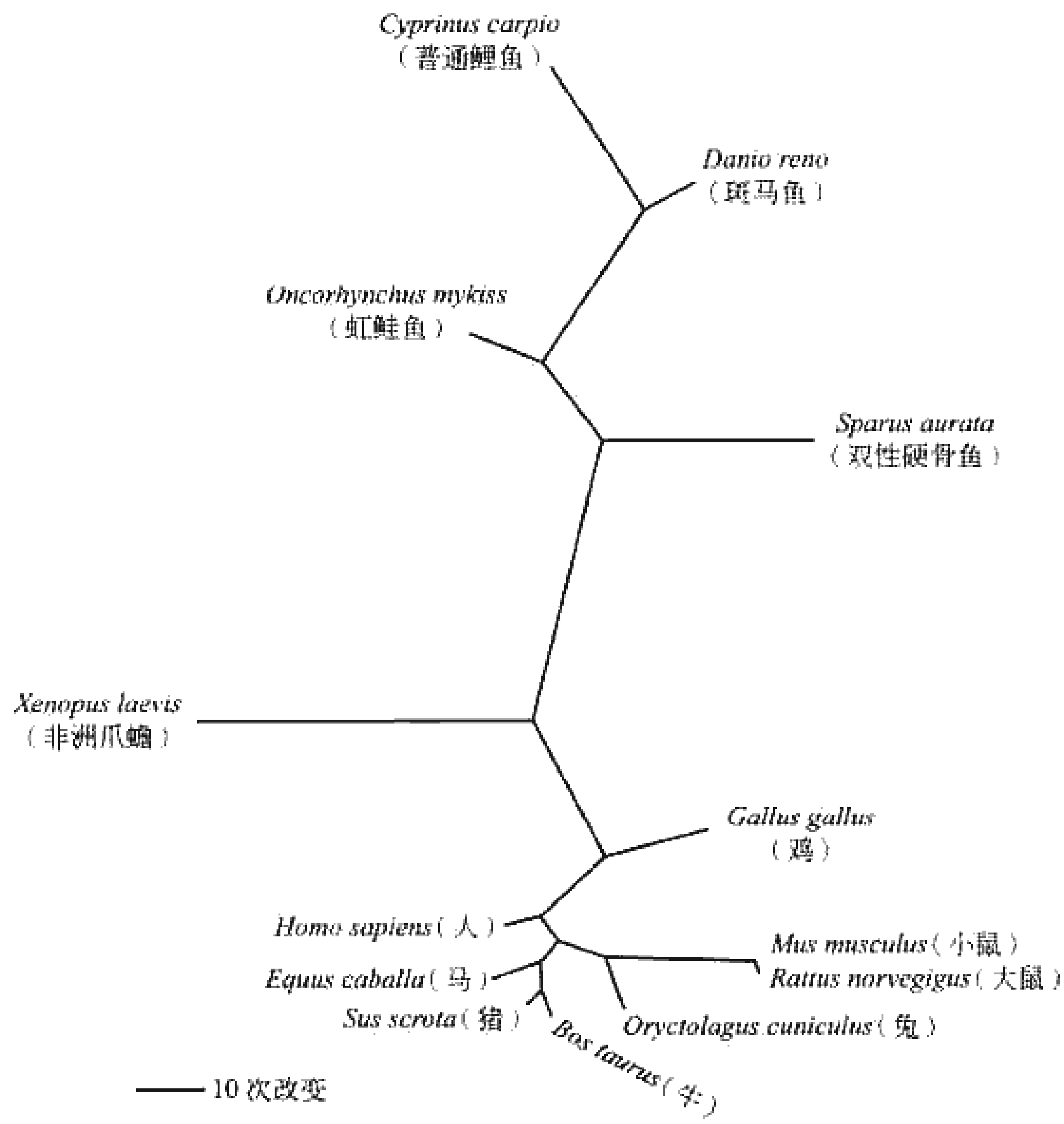


图 3-2 直系同源的 RBP 蛋白。这棵无根树是由 PAUP 软件包（见第 11 章）多序列比对 13 条 lipocalin 蛋白序列后产生得到。这些蛋白质的 accession number 为：*Homo sapiens*（NP _ 006735），*Equus caballa*（I46257），*Sus scrofa*（A39486），*Bos taurus*（P18902），*Oryctolagus cuniculus*（P06912），*Mus musculus*（Q00724），*Rattus norvegicus*（P04916），*Gallus gallus*（P41263），*Oncorhynchus mykiss*（P24744），*Xenopus laevis*（P06172），*Danio rerio*（CAB64947），*Cyprinus carpio*（CAC12738），*Sparus aurata*（AAF79012）。这棵树中，相关性越强的序列聚得越近。随着越来越多基因组被测序（第 12～第 17 章），许多家族的直系同源蛋白的数目会变得更多

人 apolipoprotein D（载脂蛋白）（NP _ 001638）具有同源性（图 3-3）。这两个蛋白是旁系同源蛋白。各 lipocalin 蛋白在体内有不同的分布，它们作为载运蛋白具有不同但是相关的功能。

Walter M. Fitch（1970）定义了这些术语。他写道："同源性中应存在两个子类。当同源是基因复制的结果，两份拷贝在一个物种的历史上是平行演化的（如 α 血红素和 β 血红素），这样的基因应被称做旁系同源基因。当同源是物种形成的结果，基因的历史反映了物种的历史（人和鼠的 α 血红素），这样的基因应被称做直系同源基因。"

我们因此定义处于同一物种内的同源基因为旁系同源基因。但是考察球蛋白的例子时，人 α 球蛋白和 β 球蛋白是旁系同源的，鼠 α 球蛋白和 β 球蛋白也是如此。人 α 球蛋白和鼠 α 球蛋白是直系同源的。人 α 球蛋白和鼠 β 球蛋白是什么关系呢？它们能被认为是旁系同源关系，因为 α 球蛋白和 β 球蛋白产生于基因复制事件而不是物种形成事件。但是它们不是旁系同源，因为它们不在同一物种内，因此简单认为它们"同源"比较合适，反映了它们来自同一祖先基因。Fitch（1970）提出研究种系发生时也需要研究直系同源物（见第 11 章）。

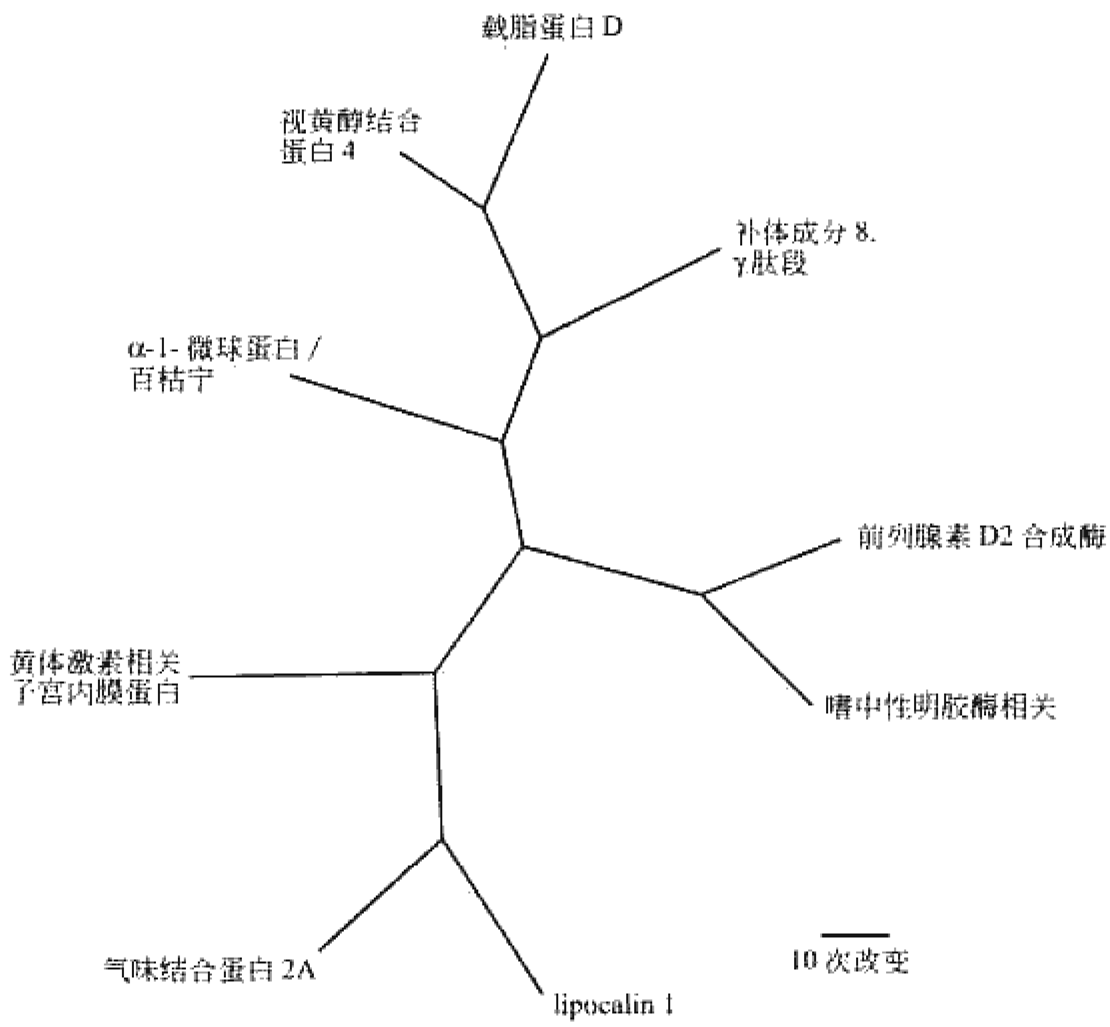


图 3-3 人 lipocalin 旁系同源蛋白：每个蛋白都是人蛋白，每个蛋白都属于 lipocalin 家族。这棵无根树由 PAUP 程序（见第 11 章）产生。这些蛋白质的 accession number 为：α-1-微球蛋白/bikunin（NP _ 001624），载脂蛋白 D（NP _ 001638），补体成分 8，补体成分 γ（NP _ 000597），lipocalin 1（NP _ 002288），嗜中性白明胶酶相关 lipocalin（JC2339），气味结合蛋白 2A（NP _ 055397），黄体激素相关子宫内膜蛋白（XP _ 005360），前列腺素 D2 合成酶（NP _ 000945），RBP4（NP _ 006735）

Richard Owen（1804～1892）是第一位使用"同源"这个词的生物学家。他定义"同源"为"不同动物中在每个形式和功能上都相同的器官"（Owen，1843）。

值得一提的是，直系同源物不一定需要有相同的功能。在第 8 章我们将给出基因和蛋白质功能的各种定义。之后，我们会沿着生命之树探索基因组（第 12～第 17 章）。在所有的基因组测序计划中，直系同源物是通过数据库搜索确定的。如果两条 DNA（蛋白质）序列相互比对的得分较高，那么它们就被定义为同源 DNA（蛋白质），这将在第 4 章讨论。但是，同源蛋白不必要有相同功能。

我们可以通过双序列比对评估任意两条蛋白之间的相关性。在这个过程中，我们把两条序列紧挨着放在一起。RBP 和 β-乳球蛋白的序列如图 3-4 所示。我们用单字符氨基酸编码，并缩写列于框 3-1 中。

人 RBP (199 个氨基酸；访问编号为 NP_006735)

```
1      MKWVWALLLL AAWAAAERDC RVSSFRVKEN FDKARFSGTW YAMAKKDPEG
51     LFLQDNIVAE FSVDETGQMS ATAKGRVRLL NNWDVCADMV GTFTDTEDPA
101    KFKMKYWGVA SFLQKGNDDH WIVDTDYDTY AVQYSCRLLN LDGTCADSYS
151    FVFSRDPNGL PPEAQKIVRQ RQEELCLARQ YRLIVHNGYC DGRSERNLL
```

牛 β-乳球蛋白 (178 个氨基酸；访问编号为 P02754)

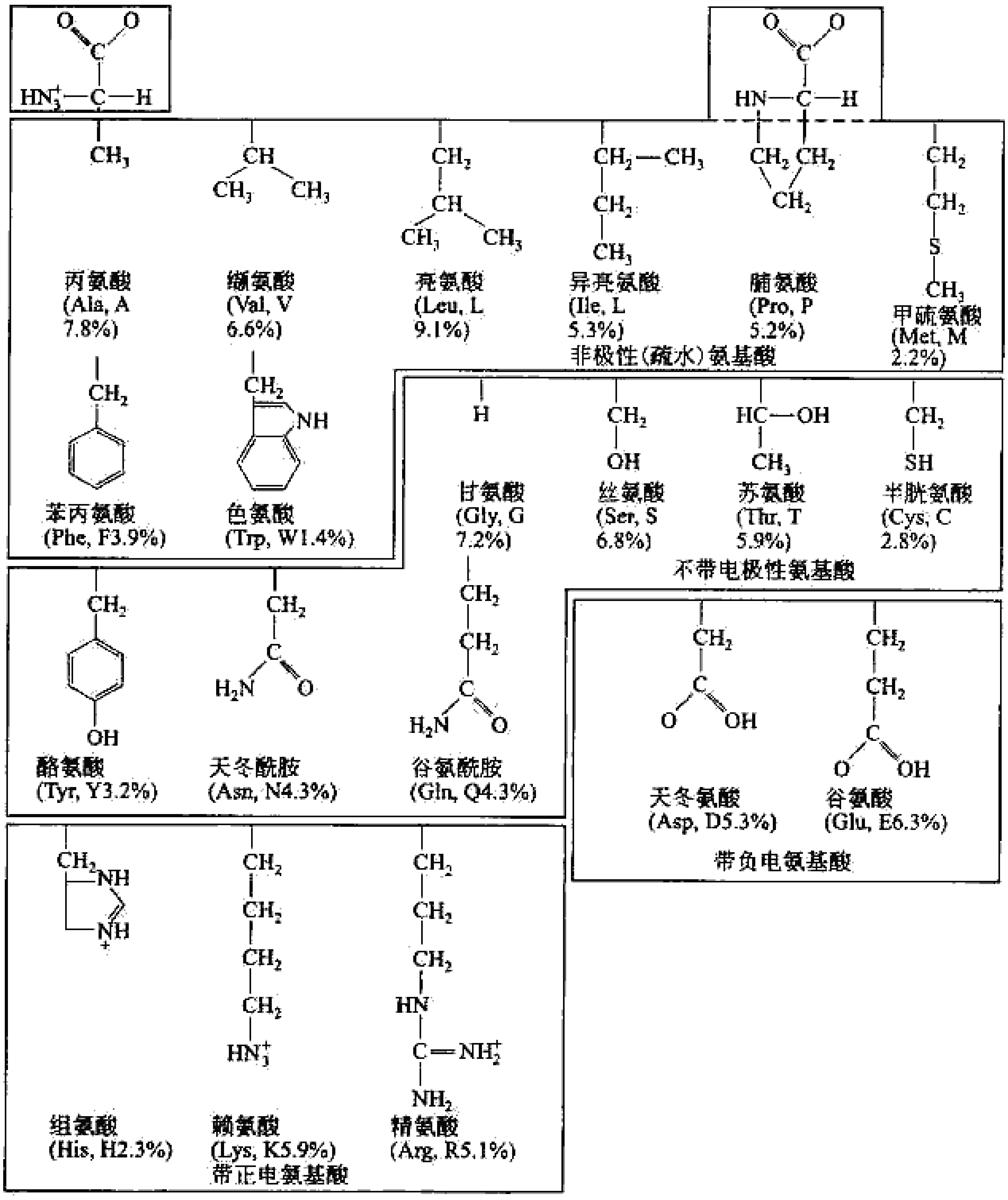
```
1      MKCLLLALAL TCGAQALIVT QTMKGLDIQK VAGTWYSLAM AASDISLLDA
51     QSAPLRVYVE ELKPTPEGDL EILLQKWENG ECAQKKIIAE KTKIPAVFKI
101    DALNENKVLV LDTDYKKYLL FCMENSAEPE QSLACQCLVR TPEVDDEALE
151    KFDKALKALP MHIRLSFNPT QLEEQCHI
```

图 3-4 RBP 和 β-乳球蛋白的序列（以单字符氨基酸编码表示）

通过视觉检查来比对这两条序列是十分困难的。而且，如果我们允许比对中有空隙出现以表示序列中出现的删除和插入，那么可能的比对的数目会呈指数增长。显然，我们需要一计算机程序来执行这个比对（框 3-2）。两个蛋白的双序列比对在图 3-5 中有所阐述。在这个比对中，RBP 位于上部，β-乳球蛋白在下部。竖线（|）代表比对中出现相同的氨基酸。

框 3-1　20 个常见氨基酸的结构和单字符、三字符缩写

记住这些字符的缩写并熟悉氨基酸的物理特性非常有帮助。百分数代表了各氨基酸在蛋白质中的相对丰度。



框 3-2　算法和程序

算法是由计算机程序组织起来的过程（Sedgewick，1988）。例如，许多算法可用于双序列比对。计算机程序是一些指令的集合，它使用一个或多个算法来解决一项任务。例如，BLAST 程序（第 3～第 5 章）使用一系列算法来执行序列比对。在第 11 章将会介绍其他程序使用算法来产生系统发生树。

要解决一系列的生物信息学问题，计算机程序是必需的，因为这些问题需要成千上万次操作。程序所使用的算法决定了这些操作是如何被自动化的。贯穿全书，我们会提到许多程序，它们使用了各种不同的算法。每一个程序和算法都是用来解决一项特定的任务。用于一条蛋白质序列和另一条序列比较的程序不能用于比较一条序列和由千万条蛋白质序列组成的数据库。

为什么用于比较两条序列有用的算法不能用于比较上万条序列？有些问题固有的复杂性使得穷尽分析方法需要大量的计算机内存或者需要难以接受的长时间才能完成这些任务。一个启发式算法对最佳答案采取近似策略，并不需要穷尽考虑每一步可能的结果。图 3-2 中的 13 条蛋白组成一棵树的不同方式有 10 亿种以上（见第 11 章）一启发式方法能在 1s 内解决这个找到最优树的问题。

```
                     .         .         .         .         .
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
       .   |||  |       .    |.   .   .   |  : .||||.:|      :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

                     .         .         .         .         .
 51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
    : | |    |    |       ::  | .| .   ||   |:    ||           |.
 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

                     .         .         .         .         .
 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV..........QYSC 136 RBP
     || ||.         |          :.|||  | .            .|
 94 IPAVFKIDALNENKVL........VLDTDYKKYLLFCMENSAEPEQSLAC 135 lactoglobulin

                     .         .         .         .         .
137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
     . |       |      | : |  ||    .      | || |
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI....... 178 lactoglobulin
```

图 3-5 人RBP和牛β-乳球蛋白的双序列比对。注意比对是全局的（即蛋白质序列的全部长度都参与比较），两条序列间有许多相同的位置（由 | 表示）。这个比对中有五种不同的点。①比对残基之间的两点表示不同程度的相似性（例如在顶部R和K之间有两点，它们有相似的物理化学性质）（见箭头1）。②比对残基之间的单点（例如箭头2）也表示相似性，但相似程度比两点少。③，④比对包含的内部间隙（由沿着序列字符位置处的点表示，例如箭头3）和β-乳球蛋白氨基端和羧基端的端部间隙（见箭头4）。⑤序列上方的点每10个位置出现一次（箭头5）

图3-5的比对是由Genetics Computer Group (GCG)软件包的GAP程序所生成。在本章的后面，我们将介绍用于双序列比对的基于Web的免费工具和商业工具包（如GCG）。

本章的后面我们将会知道相似的一对残基在用于比对打分的取代矩阵（substitution matrix）中会有一个正的分值。

例如，在最上部的一行中，残基GTWY在两条序列中是相同的。我们可以数一下相同残基的数目，在这个例子中，两条序列有26%的一致性（43个残基）。一致性表示了两氨基酸（核苷酸）序列相同的程度。

双序列比对的另一方面是一些比对的残基相似但不相同。它们相关，因为它们有一些相同的生物化学性质。相似的残基对在结构和功能上相关。例如，在比对的第一行我们能发现精氨酸和赖氨酸（R和K以两点相连）；除此我们也会发现天冬氨酸和谷氨酸残基比对在一起。这些是保守性替代（conservative substitution）。具有相似属性的氨基酸包括碱性氨基酸（K，R，H）、酸性氨基酸（D，E）、羟基化氨基酸（S，T）和疏水性氨基酸（W，F，Y，L，I，V，M，A）。

两条蛋白质序列的相似百分比（percent similarity）是相等和相似匹配之和。在图3-5的最上部分共有44个比对的氨基酸残基，其中11个相等，3个相似。相同百分比（percent identity）是25%（11/44），而相似百分比是32%（14/44）。一般地，考虑两条蛋白质序列之间的一致程度比相似程度更有用，因为相似性的衡量依赖于如何定义两个氨基酸残基是如何相关（相似）的。我们将在第4章看到使用BLAST搜索数据库时给出的双序列比对中有相同百分比和"正量（positives）"（相似百分比）（图4-6）。

概而言之，双序列比对是排列两条序列以达到最大程度相同（在氨基酸比对的情况下是

最大程度的保守性）的过程。双序列比对的目的是衡量两个分子的相似性和同源的可能性。我们可以说两个蛋白有25%的同一性或者（上面的比对）32%的相似性。如果序列相同的程度较为显著，那么两条序列有可能同源。说两条序列有一定程度的同源性是不对的，因为它们要么同源要么不同源。我们下面将要讨论序列比对的统计显著性，这些分析为衡量两个蛋白同源这个假设提供了证据。最终两条蛋白是否同源的最强的证据来自于结构研究和进化分析。

由于趋同进化（convergent evolution），两蛋白也可能有相似结构，因此分子进化研究是重要的。

3.1.3 间隙

双序列比对可以作为一个有用的工具来确定那些在进化过程中发生的并且导致我们所研究的两条蛋白质序列产生分歧的突变。常见的突变包括替换、插入和删除。替换发生于一个突变导致一种氨基酸的密码子变成另一种氨基酸的密码子时。这在比对中的结果是在同一位置上出现两个不相同的氨基酸，例如丝氨酸和苏氨酸。插入和删除发生于残基添加或消除时，比对中由一无效字符表示（句点号 .），加在一条或另一条序列中。插入或删除（即使只有一个字符长度）都被认为是比对中的间隙。

在我们对人RBP4和牛β-乳球蛋白的比对中，有8处间隙。两处发生在蛋白质的端部，较长的两处发生在中间，其他四处间隙只有一个或两个氨基酸。注意间隙的效果是使得两个排列在总体长度上相等。

比对中间隙的加入可能是生物学相关的，因为间隙反映了进化上发生的变化。在应用上，间隙使得两个蛋白质能进行全长比对。

3.1.4 双序列比对、同源和生命的进化

如果两条蛋白是同源的，那么它们有一个共同的祖先。通常，我们从物种中获得的蛋白质（或基因）序列是现存的序列。我们可以比较来自不同物种的 *RBP* 序列（如人、鱼和虹鳟鱼），然后发现这些序列是同源的（图3-2）。这意味着存在一祖先物种，它拥有 *RBP* 基因，生存于人和虹鳟鱼种系分化以前（大约400MYA）。这个祖先生物的后代包括除人和虹鳟鱼之外的许多物种。使用双序列比对对蛋白质（或DNA）序列的同源研究也有助于调查蛋白质（或基因）的进化历史。

推断出共同祖先的序列是有可能的（见第11章）。

图3-6为地球上生命进化的时间尺度的简要概览（详细讨论参见第12章）。不同物种的分化是通过使用多种来源的数据建立的，特别是化石记录。原核生物的化石发现于35亿年前的岩石中。在lipocalin的例子中，没有在无脊椎动物（如昆虫）中发现 *RBP* 的直系同源，但是在许多鱼类和两栖类中找到 *RBP*（图3-2）。我们因此能推测 *RBP* 基因起源于400～700MYA之间。其他的lipocalin基因更古老。例如，细菌lipocalin基因大概产生于2BYA。

当我们研究一系列同源蛋白质序列时，我们可以发现这些家族的成员有很大的保守性，一些非常古老的成员也相当保守。例如甘油醛-3-磷酸脱氢酶（GAPDH）。多序列比对，实质上是一组蛋白质之间的一系列的双序列比对，揭示了GAPDH直系同源是相当保守的（图3-7）。

来自不同物种的直系同源RBP提供了另一个高保守性家族的例子（图3-8）。比对中有

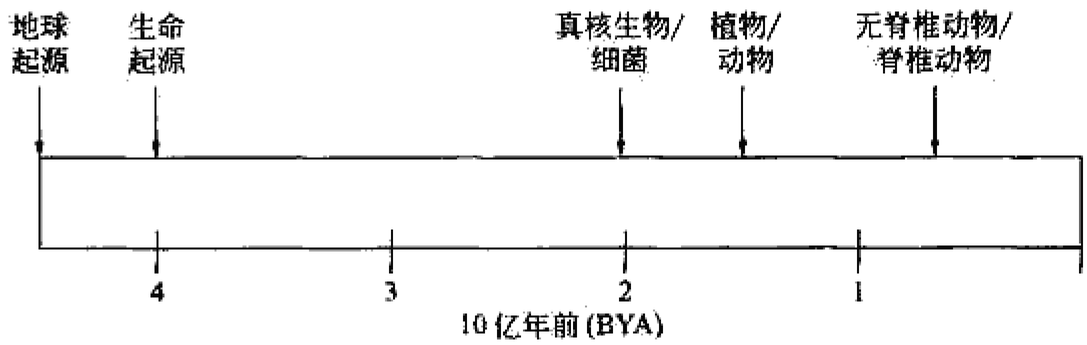


图 3-6 地球上生命历史纵览（详细参见第 12 章）。在进化的背景下研究基因/蛋白质序列：哪些生物有直系同源基因？这些生物什么时候进化？人和细菌 lipocalin 如何相关

```
果蝇   GAKKVIISAP SAD.APM..F VCGVNLDAYK PDMKVVSNAS CTTNCLAPLA
人     GAKRVIISAP SAD.APM..F VMGVNHEKYD NSLKIISNAS CTTNCLAPLA
植物   GAKKVIISAP SAD.APM..F VVGVNEHTYQ PNMDIVSNAS CTTNCLAPLA
细菌   GAKKVVMTGP SKDNTPM..F VKGANFDKY. AGQDIVSNAS CTTNCLAPLA
酵母   GAKKVVITAP SS.TAPM..F VMGVNEEKYT SDLKIVSNAS CTTNCLAPLA
古细菌 GADKVLISAP PKGDEPVKQL VYGVNHDEYD GE.DVVSNAS CTTNSITPVA

果蝇   KVINDNFEIV EGLMTTVHAT TATQKTVDGP SGKLWRDGRG AAQNIIPAST
人     KVIHDNFGIV EGLMTTVHAI TATQKTVDGP SGKLWRDGRG ALQNIIPAST
植物   KVVHEEFGIL EGLMTTVHAT TATQKTVDGP SMKDWRGGRG ASQNIIPSST
细菌   KVINDNFGII EGLMTTVHAT TATQKTVDGP SHKDWRGGRG ASQNIIPSST
酵母   KVINDAFGIE EGLMTTVHSL TATQKTVDGP SHKDWRGGRT ASGNIIPSST
古细菌 KVLDEEFGIN AGQLTTVHAY TGSQNLMDGP NGKP.RRRRA AAENIIPTST

果蝇   GAAKAVGKVI PALNGKLTGM AFRVPTPNVS VVDLTVRLGK GASYDEIKAK
人     GAAKAVGKVI PELNGKLTGM AFRVPTANVS VVDLTCRLEK PAKYDDIKKV
植物   GAAKAVGKVL PELNGKLTGM AFRVPTSNVS VVDLTCRLEK GASYEDVKAA
细菌   GAAKAVGKVL PELNGKLTGM AFRVPTPNVS VVDLTVRLEK AATYEQIKAA
酵母   GAAKAVGKVL PELQGKLTGM AFRVPTVDVS VVDLTVKLNK ETTYDEIKKV
古细菌 GAAQAATEVL PELEGKLDGM AIRVPVPNGS ITEFVVDLDD DVTESDVNAA
```

图 3-7 来自 6 种生物的甘油醛-3-磷酸脱氢酶部分多序列比对。*Drosophila melanogaster*（果蝇，GenBank accession P07487），*H. sapiens*（NP _ 002037），*Arabidopsis thaliana*（一种植物，AAD30223），*Escherichia coli*（大肠杆菌，P06977），*Saccharomyces cerevisiae*（酵母，NP _ 011708），以及 *Haloarcula vallismortis*（古细菌）（Q48335）。比对由 GCG 中的 PileUp 程序完成（见第 10 章和附录）。注意到古细菌序列分歧最大，但总体上这些直系同源序列相当保守（古细菌是微生物，它形成生物三界之一，见第 12 章和第 14 章）

```
                                           ↓ ↓         ↓    ↓
(1)  VKQDFDRMRYQGTW YAVAKKDPVG LFLLDNVVAN FKVQEDGTMT
(2)  VQQDFNRTRYQGTW YAVAKKDPVG LFLLDNIVAN FKVEEDGTMT
(3)  VMQNFDRSRYTGRW YAVAKKDPVG LFLLDNVVAQ FSVDESGKVT
(4)  VMQNFDKTRYAGTW YAVGKKDPEG LFLIDNIVAQ FTIHEDGAMT
(5)  VKENFDKARFSGLW YAIAKKDPEG LFLQDNIIAE FSVDEKGHMS
(6)  VKENFDKARFSGLW YAIAKKDPEG LFLQDNIIAE FSVDEKGHMS
(7)  VKENFDKARFAGTW YAMAKKDPEG LFLQDNIVAE FSVDENGHMS
(8)  VKENFDKARFSGTW YAMAKKDPEG LFLQDNIVAE FSVDENGHMS
(9)  VKENFDKARFSGTW YAMAKKDPEG LFLQDNIVAE FSVDEYGQMS
(10) VKENFDKARFSGTW YAMAKKDPEG LFLQDNIVAE FSVDETGQMS
(11) VKENFDKARFAGTW YAMAKKDPEG LFLQDNIVAE FSVDENGHMS
(12) VKENFDKNRYSGTW YAMAKKDPEG LFLQDNVVAQ FTVDENGQMS
(13) VMKDFNKERYAGVW YAVAKKDPEG LFLLDNIAAN FKIEDNGKTT
```

图 3-8 RBP 多序列比对的部分。注意许多残基十分保守（包括 GXW motif，以阴影表示）。Accession Number 在图 3-2 的旁注中已给出。蛋白质来自（1）鲤鱼（*C. carpio*）；（2）斑马鱼（*D. rerio*）；（3）虹鳟鱼（*O. mykiss*）；（4）金头鲷（*S. aurata*）；（5）小鼠（*M. musculus*）；（6）大鼠（*R. norvegicus*）；（7）牛（*B. taurus*）；（8）猪（*S. scrofa*）；（9）马（*E. caballa*）；（10）人（*H. sapiens*）；（11）兔（*O. cuniculus*）；（12）鸡（*G. gallus*）以及（13）蛙（*X. laevis*）

多列高度保守，包括甘氨酸-X-色氨酸（GXW，X 指任何氨基酸）模体，这是数百 lipocalin 蛋白的特点。注意到一些位置的氨基酸发生保守性替换，例如比对的最后位置上丝氨酸和苏氨酸的替换（图 3-8）。另一些位置的保守性小一些，在紧挨经典 GXW 模体的前面，氨基酸可以是谷氨酰胺、苏氨酸、丝氨酸或者丙氨酸。一些构成视黄醇结合口袋的氨基酸残基也相当保守（Newcomer 等，1984）（图 3-8，箭头所示）。

我们以人 lipocalin 家族作为多序列比对的最后一个例子。这个家族的成员是旁系同源物，但是高度趋异（图 3-9）。比对中仅存的一个保守性模体是经典的 GXW 模体。尽管氨基酸序列的高度趋异，该家族的所有成员都采用了十分相似的三维结构（见图 3-1 和第 9 章）。

```
(1) ~~~~~EIQDVSGTWYAMTVDREFPEMNLESVTPMTLTTL.GGNLEAKVTM
(2) LSFTLEEEDITGTWYAMVVDKDFPEDRRRKVSPVKVTALGGGNLEATFTF
(3) TKQDLELPKLAGTWHSMAMATNNISLMATLKAPLRVHITSEDNLEIVLHR
(4) VQENFDVNKYLGRWYEIEKIPTTFENGRCIQANYSLMENGNQELRADGTV
(5) VKENFDKARFSGTWYAMAKDPEGLFLQDNIVAEFSVDETGNWDVCADGTF
(6) LQQNFQDNQFQGKWYVVGLAGNAI.LREDKDPQKMYATIDKSYNVTSVLF
(7) VQPNFQQDKFLGRWFSAGLASNSSWLREKKAALSMCKSVDGGLNLTSTFL
(8) VQENFNISRIYGKWYNLAIGSTCPWMDRMTVSTLVLGEGEAEISMTSTRW
(9) PKANFDAQQFAGTWLLVAVGSACRFLQRAEATTLHVAPQGSTFRKLD...
```

图 3-9 人 lipocalin 旁系同源物多序列比对的部分。请注意保守的 GXW 模体（方括号内）。所有序列都来自于 *H. sapiens*。这些蛋白质是（1）lipocalin 1；（2）气味结合蛋白 2A；（3）孕激素相关子宫内膜蛋白；（4）载脂蛋白 D；（5）RBP4；（6）嗜中性明胶酶相关蛋白；（7）前列腺素 D2 合成酶；（8）α-1-微球蛋白/bikunin；以及（9）补体 8（γ 肽）。accession number 在图 3-3 的旁注中给出。请注意我们除去了比对中出现间隙的许多列

从上面的例子我们可以看到任意两个蛋白之间的双序列比对呈现出不同程度的保守性。Margaret Dayhoff（1978）提出了一个衡量蛋白质进化变化的模型。我们将考察 Dayhoff 模型，它提供了双序列比对打分系统的基础。这个系统给任意两个蛋白质打分，衡量它们或近或远的相关性。然后我们将讨论两种双序列比对算法，包括全局的和局部的。许多数据库搜索方法（如 BLAST）（第 4 章和第 5 章）都以某种形式建立在 Dayhoff 模型对进化的分析之上。

Dayhoff（1978）的这篇文献是指 *Atlas of Protein Sequence and Structure*，它是一本共 25 章的书（有多个作者），描述了蛋白质家族。

3.2 Dayhoff 模型：可接受点突变

Dayhoff 和同事整理了数千种蛋白质并且把许多家族中紧密相关的蛋白质做了比较。他们考虑这样的问题：当两个同源蛋白比对时，我们能观察到哪些氨基酸被替代？他们定义可接受点突变（accepted point mutation）为在蛋白质中被自然选择接受的单个氨基酸替换。可接受点突变简称为 PAM（因为它比 APM 容易发音）。被自然选择接受的氨基酸变化发生于：（1）一条基因发生 DNA 突变，该突变导致编码另一个不同的氨基酸；（2）整个物种都认可这种变化，把它作为蛋白质的主要形式。

Dayhoff 和同事没有比较一个残基直接突变成另一个残基的不同概率。实际上，他们使用简约分析法（parsimony analysis）来构建系统发生树（见第 11 章）。然后，他们描述了两个比对在一起的残基来自于同一个祖先残基的概率。通过这种方法，他们就把比对的残基对中的多取代复杂效应最小化了。

蛋白质进化中哪些点突变是可以接受的？直观上，保守性取代，例如丝氨酸取代苏氨酸最有可能被接受。为了考察所有的变化，Dayhoff和其同事研究了71组紧密相关的蛋白质中1572个变化（框3-3）。因此，他们对"可接受"突变的定义是基于经验观察的。结果显示于图3-10中，描述了任一氨基酸对（i,j）比对在一起的频率。今天，我们能使用多得多的数据来产生这样的表（参见图2-1和GenBank数据的爆炸式增长）。Gonnet和合作者提供了PAM矩阵的更新版本（Gonnet等，1992）。虽然如此，1978年发现的版本是基本正确的。一些氨基酸残基（如天冬酰胺和丝氨酸）产生替代较为频繁，其他残基（特别是色氨酸和半胱氨酸）则很少发生突变。

我们将在第8章中讨论蛋白质超家族

框3-3 Dayhoff的蛋白质超家族

Dayhoff（1978）研究了34个蛋白质"超家族"，并把它们聚成71个系统发生树。这些蛋白质包括从高度保守的蛋白质（如组蛋白和谷氨酸脱氢酶）到那些有高可接受突变率的蛋白质（如免疫球蛋白链和载体蛋白）。比对蛋白质家族（比较图3-7），然后他们衡量了比对中一个氨基酸被另一个氨基酸取代的概率。这里列出了部分他们研究的蛋白质，包括突变接受率。详细信息参见表3-1。进化最快和最慢的家族其速率相差近40倍，但是在同一家族内，进化的速率（由单位时间内的PAM数来衡量）不同物种之间仅有2～3倍的差别。此表被授权使用。

表3-1 不同蛋白质的进化速率

蛋白质	每亿年的PAM数	蛋白质	每亿年的PAM数
免疫球蛋白(Ig)κ链C区	37	胰蛋白酶	5.9
κ酪蛋白	33	胰岛素	4.4
上皮生长因子	26	细胞色素c	2.2
血清白蛋白	19	谷氨酸脱氢酶	0.9
血红蛋白α链	12	组蛋白H3	0.14
肌球蛋白	8.9	组蛋白H4	0.10
神经生长因子	8.5		

Dayhoff等计算了各氨基酸的相对突变能力（表3-2）。这只是描述了在较短进化时期内每个氨基酸发生变化的频度（我们刚在描述中指出的进化时间是短的，因为参与分析的蛋白质是紧密相关的）。为了计算相对突变能力，他们把每种氨基酸发生突变的次数除以该氨基酸出现的总次数。表3-3显示了每种氨基酸的出现频率。

表3-2 氨基酸相对突变可能性

天冬酰胺	134	组氨酸	66
丝氨酸	120	精氨酸	65
天冬氨酸	106	赖氨酸	56
谷氨酸	102	脯氨酸	56
丙氨酸	100	甘氨酸	49
苏氨酸	97	酪氨酸	41
异亮氨酸	96	苯丙氨酸	41
甲硫氨酸	94	亮氨酸	40
谷氨酰胺	93	半胱氨酸	20
缬氨酸	74	色氨酸	18

注：1. 丙氨酸的值设为100。
2. 来源于Dayhoff（1978），并被授权使用。

表3-3 氨基酸归一化频率

甘氨酸	0.089	精氨酸	0.041
丙氨酸	0.087	天冬酰胺	0.040
亮氨酸	0.085	苯丙氨酸	0.040
赖氨酸	0.081	谷氨酰胺	0.038
丝氨酸	0.070	异亮氨酸	0.037
缬氨酸	0.065	组氨酸	0.034
苏氨酸	0.058	半胱氨酸	0.033
脯氨酸	0.051	酪氨酸	0.030
谷氨酸	0.050	甲硫氨酸	0.015
天冬氨酸	0.047	色氨酸	0.010

注：1. 所有值和为1。如果20种氨基酸在蛋白质中等概率出现，这些值都是0.05（即5%）。事实上，氨基酸出现的频率不同。
2. 来源于Dayhoff（1978），并被授权使用。

原氨基酸

替代氨基酸

	A 丙氨酸	R 精氨酸	N 天冬酰胺	D 天冬氨酸	C 半胱氨酸	Q 谷氨酰胺	E 谷氨酸	G 甘氨酸	H 组氨酸	I 异亮氨酸	L 亮氨酸	K 赖氨酸	M 甲硫氨酸	F 苯丙氨酸	P 脯氨酸	S 丝氨酸	T 苏氨酸	W 色氨酸	Y 酪氨酸	V 缬氨酸
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	0	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	A 丙氨酸	R 精氨酸	N 天冬酰胺	D 天冬氨酸	C 半胱氨酸	Q 谷氨酰胺	E 谷氨酸	G 甘氨酸	H 组氨酸	I 异亮氨酸	L 亮氨酸	K 赖氨酸	M 甲硫氨酸	F 苯丙氨酸	P 脯氨酸	S 丝氨酸	T 苏氨酸	W 色氨酸	Y 酪氨酸	V 缬氨酸

图 3-10 可接受点突变数目（乘以 10，来自于紧密相关蛋白质序列集中的 1572 个氨基酸替换）。该图来自 Dayhoff（1978）并做了修改。氨基酸以其三字母编码的字符顺序排列。注意一些替换经常被接受（如 V 和 I 或者 S 和 T）。其他氨基酸（如 C 和 W）很少被其他氨基酸替换。此图被授权使用

为什么有些氨基酸比另一些更容易发生突变？不容易发生突变的残基可能对蛋白质的结构和功能较重要，这样以其他氨基酸来替代的序列对生物有害（在第18章我们将会看到许多人类疾病，从囊肿纤维症到与孤独症相关的Rett综合征，都是由某个蛋白质中的单个氨基酸替换导致的）。相反地，最容易突变的氨基酸（天冬酰胺、丝氨酸、天冬氨酸和谷氨酸）在蛋白质中的功能很容易被其他残基取代。最常发生的替换（图3-10）是谷氨酸对天冬氨酸（都是酸性氨基酸）、丝氨酸对丙氨酸、丝氨酸对苏氨酸（都是羟基化氨基酸）和异亮氨酸对缬氨酸（都是疏水性氨基酸而且有相同大小）。

Dayhoff等研究的蛋白质有85%以上的一致性。因此，他们能构建具有高可信度的比对。本章的后面我们将会看到Needleman和Wunsch算法（1970）可以自动比对蛋白质序列。

蛋白质中发生的替换也能从遗传密码的角度来理解（框3-4）。请观察氨基酸发生替换仅需一个碱基发生改变的概率有多大。例如，天冬氨酸由GAU或GAC编码，第三个位置变成A或G导致密码子编码谷氨酸。也请注意5个最不容易突变的氨基酸中的4个（色氨酸、半胱氨酸、苯丙氨酸和酪氨酸）仅有两个或一个密码子。编码W（色氨酸）的三个碱基中突变任何一个都能导致氨基酸的变化。这些氨基酸的低突变能力意味着它们的替代突变是自然选择所不容许的。8个不容易突变的氨基酸（表3-2）中，仅有一个（亮氨酸）有6个密码子，两个（甘氨酸和脯氨酸）有4个密码子。其他有一个或两个密码子。

每一对氨基酸发生替代的观察频率称为"目标频率"。它们构成了我们在这一节叙述的取代矩阵的基础。

你可以使用欧洲生物信息学研究所（EBI）提供的蛋白质组工具（http://www.ebi.ac.uk/proteome/）来查看每个氨基酸的发现频率。在这个网站中点击 *Homo sapiens*，然后点击"primary amino acid composition"。

框3-4 遗传密码

UUU	171	F	UCU	147	S	UAU	124	Y	UGU	99	C
UUC	203	F	UCC	172	S	UAC	158	Y	UGC	119	C
UUA	73	L	UCA	118	S	UAA	0	Stop	UGA	0	Stop
UUG	125	L	UCG	45	S	UAG	0	Stop	UGG	122	W
CUU	127	L	CCU	175	P	CAU	104	H	CGU	47	R
CUC	187	L	CCC	197	P	CAC	147	H	CGC	107	R
CUA	69	L	CCA	170	P	CAA	121	Q	CGA	63	R
CUG	392	L	CCG	69	P	CAG	343	Q	CGG	115	R
AUU	165	I	ACU	131	T	AAU	174	N	AGU	121	S
AUC	218	I	ACC	192	T	AAC	199	N	AGC	191	S
AUA	71	I	ACA	150	T	AAA	248	K	AGA	113	R
AUG	221	M	ACG	63	T	AAG	331	K	AGG	110	R
GUU	111	V	GCU	185	A	GAU	230	D	GGU	112	G
GUC	146	V	GCC	282	A	GAC	262	D	GGC	230	G
GUA	72	V	GCA	160	A	GAA	301	E	GGA	168	G
GUG	288	V	GCG	74	A	GAG	404	E	GGG	160	G

表中描述了64种可能的密码子和这些密码子使用的频率以及它们对应的氨基酸单字母。总共有4种碱基（A、C、G、U），每密码子有3个碱基，所以共有$4\times4\times4=4^3=64$种密码子。

请注意遗传密码的一些特性。表示一种氨基酸的密码子可能有 1 个（M、W），2 个（C、D、E、F、H、K、N、Q、Y），3 个（I），4 个（A、G、P、T、V）或者 6 个（L、R、S）。很少情况下 UGA 被读作硒代半胱氨酸（selenocysteine）（简写为 Sec，单字母简写是 U）。

表中每一区块里排列在一起的 4 个密码子，有一个经常使用得很少。例如，对于 F、L、I、M 和 V（即中间是 U 的密码子，位于遗传密码表的第一列），第三个位置上腺嘌呤（A）使用得相对较少。对于中间是胞嘧啶的密码子，第三个位置是鸟嘌呤的机会就很低。

也请注意许多情况下，突变导致的是氨基酸的保守性改变（或者根本不变）。苏氨酸（ACX）第三个位置的任何变化都不会改变氨基酸，这是“摇摆（wobble）”特性。如果苏氨酸密码子的第一个核苷酸从 A 变成 U，这就保守地取代成丝氨酸；如果第二个核苷酸从 C 变成 G，也能替换成丝氨酸。类似的保守替换模式也能在第一列的密码子中找到，那里所有的残基都是疏水的，D、E、K 和 R 都是带电氨基酸。

不同生物和同一生物内的不同基因，密码子使用是不同的。也注意这是标准遗传密码，一些生物使用另外的遗传密码。

注：取自 International Human Genome Sequencing Consortium（2001）中图 34，并被授权使用。

3.2.1 PAM1 矩阵

Dayhoff 和其同事接下来用这些可接受突变的数据（图 3-10）和每个氨基酸的发现频率来产生突变概率矩阵（mutation probability matrix）M（图 3-11）。矩阵元素 M_{ij} 表示在一给定进化时期内氨基酸 j（列）替换成氨基酸 i（行）的概率。在图 3-11 的例子内，进化时期为一个 PAM［PAM 定义为进化趋异（evolutionary divergence）的单位，表示两个蛋白 1%氨基酸发生变化的时间］。注意这个 PAM 矩阵的进化期间是以氨基酸趋异的百分比来定义的，而不是以年为单位。如我们所知（比较图 3-7、图 3-8 和图 3-9），蛋白质序列 1%的趋异在以不同替换速率进化的蛋白质家族中的时间尺度是不同的。

观察图 3-11 会发现几个重要特性。最高分位于从左上至右下的对角线上。每列值的和为 10000（对应 100%）。顶部左边的值 9867 意味着在一个 PAM 的进化期间内原序列中的丙氨酸将有 98.67%的机会被替换成丙氨酸，有 0.28%的机会变成丝氨酸。最容易发生突变的氨基酸（根据表 3-2）——天冬氨酸仅有 98.22%的机会不发生改变。最不容易突变的氨基酸——色氨酸有 99.7%的概率保持不变。

对每一个原始氨基酸，很容易发现如果要发生改变，其最可能会变成哪个氨基酸。这些数据和双序列比对紧密相关，因为它们构成了打分系统（见下文）的基础，在比对中一个合理的氨基酸替代会得到奖分而不合理的则得到罚分。这些概念也和数据库搜索算法（如 BLAST）（第 4 章和第 5 章）相关，它们依赖于衡量分子序列相关性的打分规则。

几乎所有分子序列数据都来自于现存生物。我们可以推测祖先序列，这将在第 11 章中描述。但是一般地，对两比对的残基 a 和残基 b，我们不知道哪个会突变到哪个。Dayhoff 和其同事假设可接受氨基酸的突变是没有方向的，即两个方向具有相同的机会。在 PAM1 矩阵，蛋白质的紧密相关性使得祖先残基和两比对残基完全不同的可能性很小。

原氨基酸（列）；替代氨基酸（行）

	A 丙氨酸	R 精氨酸	N 天冬酰胺	D 天冬氨酸	C 半胱氨酸	Q 谷氨酰胺	E 谷氨酸	G 甘氨酸	H 组氨酸	I 异亮氨酸	L 亮氨酸	K 赖氨酸	M 甲硫氨酸	F 苯丙氨酸	P 脯氨酸	S 丝氨酸	T 苏氨酸	W 色氨酸	Y 酪氨酸	V 缬氨酸
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	2	2	3	1	2	1	2	0	0	9872	9	2	21	7	0	1	7	0	1	33
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

图 3-11 PAM1 突变概率矩阵。来自 Dayhoff（1978）。列为原氨基酸 j，行为取代氨基酸 i。授权使用

3.2.2 PAM250 和其他 PAM 矩阵

PAM1 矩阵基于紧密相关蛋白质序列的比对，这些蛋白质家族内的序列一致程度至少有 85%。我们经常对那些一致程度少于 85% 氨基酸一致性的蛋白质的相互关系感兴趣。我们可以通过为任意一致性程度的蛋白质构造概率矩阵来完成这项任务，考虑紧相关的蛋白，例如图 3-7 中的 GAPDH 蛋白。从一个残基到另一个残基的突变事件发生概率相对较低，用来比对这般紧相关的蛋白质的打分矩阵必须能反映这一点。在 PAM1 矩阵（图 3-11），一些氨基酸替代（如色氨酸到丝氨酸）是如此少，以至于在数据集中找不到一个例子。但是两个较远相关的蛋白（如在图 3-9 中所示的 lipocalin 蛋白）中替代发生较为普遍。PAM 矩阵（如 PAM100 或者 PAM250）就反映了远相关蛋白中这类氨基酸的替换。

除 PAM1 外的其他 PAM 矩阵是如何得来的? Dayhoff 等用 PAM1 矩阵乘以自身数百次，然后得到了其他 PAM 矩阵（框 3-5）。由此看出它们是从 PAM1 矩阵外推而得。

框 3-5　矩阵乘法

矩阵是数字的有序阵列，例如下面即为一矩阵：

$$\begin{bmatrix} 1 & 2 & 4 \\ 2 & 0 & -3 \\ 4 & -3 & 6 \end{bmatrix}$$

在对称矩阵（例如上面的例子）$a_{ij}=a_{ji}$，这表示所有对应的非对角元素是相等的。矩阵可以以各种方式被加、减和乘。如果第一个矩阵 M_1 的列数和第二个矩阵 M_2 的行数相等，那么两个矩阵可以相乘。下面的例子说明 M_1 和 M_2 如何相乘。

M_1 的每一行和 M_2 的每一列依次相乘：

$$M_1=\begin{bmatrix} 3 & 4 \\ 0 & 2 \end{bmatrix} \qquad M_2=\begin{bmatrix} 5 & -2 \\ 2 & 1 \end{bmatrix}$$

$$M_{12}=\begin{bmatrix} (3)(5)+(4)(2) & (3)(-2)+(4)(1) \\ (0)(5)+(2)(2) & (0)(-2)+(2)(1) \end{bmatrix}=\begin{bmatrix} 23 & -2 \\ 4 & 2 \end{bmatrix}$$

如果想自己尝试矩阵乘法，在 Matlab (Mathworks) 中输入图 3-11 中的 PAM1 突变概率矩阵，每个元素除以 10000，然后自乘 250 次。这样就可以得到图 3-13 的 PAM250 矩阵。

为了弄清楚不同 PAM 矩阵的意义，需要考虑极端情况。当 PAM=0 时，矩阵将成为单位矩阵（图 3-12），因为没有氨基酸发生变化。PAM 可以是相当大（如 PAM>2000，或者矩阵和自己相乘无穷次）。在 PAM∞矩阵中，每种氨基酸等概率出现，每行的所有值都接近于一个数值，这个数值就是氨基酸的出现频率（图 3-12）。我们在表 3-2 中列示了这些背景频率。

PAM250 矩阵是我们特别感兴趣的（图 3-13）。它由 PAM1 矩阵乘以本身 250 次产生，是 BLAST 搜索数据库的常用矩阵之一（第 4 章）。这个矩阵应用于有 20% 氨基酸一致性的进化距离的蛋白质。把这个矩阵和 PAM1 矩阵相互比较（图 3-11），发现许多信息内容丢失了。从左上到右下的对角线上的值倾向于比其他部分大，但没有 PAM1 矩阵中的幅度那么大。如何理解 PAM250 矩阵，举例来说，如果原氨基酸是丙氨酸，则有 13% 的概率第二条序列也将会是丙氨酸。事实上，这与丙氨酸替换成甘氨酸的机会几乎相等（12%）。对最不可突变的氨基酸——色氨酸和半胱氨酸，有>50% 的概率在这个进化期间保持不变。

原氨基酸

替代氨基酸

PAM0	A 丙氨酸	R 精氨酸	N 天冬酰胺	D 天冬氨酸	C 半胱氨酸	Q 谷氨酰胺	E 谷氨酸	G 甘氨酸
A	100%	0	0	0	0	0	0	0
R	0	100%	0	0	0	0	0	0
N	0	0	100%	0	0	0	0	0
D	0	0	0	100%	0	0	0	0
C	0	0	0	0	100%	0	0	0
Q	0	0	0	0	0	100%	0	0
E	0	0	0	0	0	0	100%	0
G	0	0	0	0	0	0	0	100%

原氨基酸

替代氨基酸

PAM∞	A 丙氨酸	R 精氨酸	N 天冬酰胺	D 天冬氨酸	C 半胱氨酸	Q 谷氨酰胺	E 谷氨酸	G 甘氨酸
A	8.7%	8.7%	8.7%	8.7%	8.7%	8.7%	8.7%	8.7%
R	4.1%	4.1%	4.1%	4.1%	4.1%	4.1%	4.1%	4.1%
N	4.0%	4.0%	4.0%	4.0%	4.0%	4.0%	4.0%	4.0%
D	4.7%	4.7%	4.7%	4.7%	4.7%	4.7%	4.7%	4.7%
C	3.3%	3.3%	3.3%	3.3%	3.3%	3.3%	3.3%	3.3%
Q	3.8%	3.8%	3.8%	3.8%	3.8%	3.8%	3.8%	3.8%
E	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%
G	8.9%	8.9%	8.9%	8.9%	8.9%	8.9%	8.9%	8.9%

图 3-12 PAM0 矩阵（上表）和 PAM∞矩阵（下表）部分。PAM∞矩阵（即 PAM1 矩阵自乘无穷次），每行的值收敛于归一化的取代氨基酸的频率（见表 3-3）。PAM2000 矩阵的值相似，也趋向于收敛于相同值。在 PAM2000 矩阵，相互比较的蛋白质无关。相反地，在 PAM0 中没有突变，蛋白质的所有残基都相当保守

原氨基酸

替代氨基酸

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

图 3-13 PAM250 突变概率矩阵。来自 Dayhoff（1978）。在这个进化距离内，5 个氨基酸残基中有 1 个从原氨基酸（列）到取代氨基酸（行）保持不变。本图已被授权使用

3.2.3 从突变概率矩阵到对数比值打分矩阵

我们研究 PAM 矩阵的目标是导出一打分系统，这样我们可以评价两条序列的相关性。当我们执行 BLAST 搜索（第 4 章和第 5 章）或者双序列比对时，我们使用一打分矩阵，但是它不是我们到现在为止讨论的形式。PAM250 突变概率矩阵（图 3-13）是有用的，因为它描述了远相关蛋白的氨基酸取代频率。我们接下来需要把 PAM 突变概率矩阵的元素转换成一打分矩阵，也叫对数比值矩阵（log-odds matrix）或者相关比值矩阵（relatedness odds matrix）。

对数矩阵的每个单元包含两个概率的“比值比（odds ratio）”，这两个概率描述了某一 PAM 间隔内一些氨基酸 a 变成氨基酸 b 的概率。比对 a、b 的分值 S 由下式给出：

$$S(a,b)=10\lg(M_{ab}/p_b) \tag{3-1}$$

在方程（3-1）中，构成“比值比”的两个概率值如下。（1）M_{ab}（例如，在 PAM250 矩阵的特殊情况下可来自于图 3-13）为真实比对（即突变被进化接受）下氨基酸残基 a、b 比对的概率。（2）归一化频率 p_b 代表随机情况下残基 b 比对的概率。第二项值为该位置上每个氨基酸 a、b 出现的概率之积。其值在表 3-3 中给出，称似然比或比值比，即 M_{ab}/p_b。

在此打分系统，Dayhoff 和其同事以 10 乘以以 10 为底的比值比的对数［见方程（3-1）］。使用对数在我们获得两序列的总比对结果时可以把比对残基的分值加起来（如果不取对数值，我们需要把这些值乘起来，这样计算上比较麻烦）。

PAM250 的对数比值矩阵列于图 3-14。所有的值近似到最近的整数。尝试使用方程（3-1），就可以明白突变概率矩阵（图 3-13）是如何转换成对数比值矩阵（图 3-14）的。例如，色氨酸在 PAM250 突变概率矩阵中值为 0.55，其归一化频率为 0.010，所以，

$$S(a,\text{tryptophan})=10\lg(0.55/0.010)=17.4 \tag{3-2}$$

A	2																			
R	−2	6																		
N	0	0	2																	
D	0	−1	2	4																
C	−2	−4	−4	−5	12															
Q	0	1	1	2	−5	4														
E	0	−1	1	3	−5	2	4													
G	1	−3	0	1	−3	−1	0	5												
H	−1	2	2	1	−3	3	1	−2	6											
I	−1	−2	−2	−2	−2	−2	−2	−3	−2	5										
L	−2	−3	−3	−4	−6	−2	−3	−4	−2	−2	6									
K	−1	3	1	0	−5	1	0	−2	0	−2	−3	5								
M	−1	0	−2	−3	−5	−1	−2	−3	−2	2	4	0	6							
F	−3	−4	−3	−6	−4	−5	−5	−5	−2	1	2	−5	0	9						
P	1	0	0	−1	−3	0	−1	0	0	−2	−3	−1	−2	−5	6					
S	1	0	1	0	0	−1	0	1	−1	−1	−3	0	−2	−3	1	2				
T	1	−1	0	0	−2	−1	0	0	−1	0	−2	0	−1	−3	0	1	3			
W	−6	2	−4	−7	−8	−5	−7	−7	−3	−5	−2	−3	−4	0	−6	−2	−5	17		
Y	−3	−4	−2	−4	0	−4	−4	−5	0	−1	−1	−4	−2	7	−5	−3	−3	0	10	
V	0	−2	−2	−2	−2	−2	−2	−1	−2	4	2	−2	2	−1	−1	−1	0	−6	−2	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

图 3-14 PAM250 对数比值矩阵。高 PAM 值（例如 PAM250）对比对高度趋异的序列有用。许多双序列比对、多序列比对和数据库搜索（如 BLAST）算法都允许选择某类 PAM 矩阵，如 PAM250、PAM70 和 PAM30

注意这个打分矩阵是对称的，不同于图3-13的突变概率矩阵。比较两序列时哪条先给出无关紧要。

色氨酸的分值+17代表同源蛋白比对中对应位置是两个色氨酸的机会是随机比对中出现的50倍。50从何得来？由方程(3-1)，使$S(a,b)=+17$，并且设取代概率$M_{ab}/p_b=x$，得到$+17=10\lg x$，$+1.7=\lg x$，因此$10^{1.7}=x=50$。

该值在PAM250对数比值矩阵中近似为17。PAM250矩阵的分值意味着什么？分值－10表示对应的氨基酸在正确的同源比对中比对在一起的频率是这些氨基酸随机比对在一起的频率的1/10（假设氨基酸是从以背景频率分布的氨基酸中随机挑选出来的）。分值0表示中性。色氨酸的分值＋17表示这种对应是随机双序列比对时比对发生的频率的50倍。＋2代表氨基酸取代是随机频率的1.6倍。矩阵中的最大值发生在色氨酸（17，此时色氨酸和色氨酸比对；12，此时半胱氨酸和半胱氨酸比对）和半胱氨酸（12），同时最大的罚分也和这两个氨基酸被取代有关。当两条序列被比对时，我们可以给出一个分值，这个分值是所有比对残基所得分值的简单加和。

通过比较PAM10（图3-15）矩阵和PAM250（图3-14）矩阵我们很容易发现不同PAM矩阵评价氨基酸取代是如何不同的。在PAM10矩阵，相同氨基酸残基对的值倾向于比PAM250中的大。例如，丙氨酸和丙氨酸匹配的分值各为7和2。PAM10矩阵中不匹配的罚分也比较大，例如，从天冬氨酸到精氨酸的突变是－17（PAM10）和－1(PAM250)。甚至，PAM10中为负值的取代（例如谷氨酸到天冬酰胺，－5）在PAM250中是正值（＋1）。为了理解PAM250和PAM10的相关性，考虑两个远相关的lipocalin蛋白的比对。使用PAM250，许多氨基酸替换仅得到较小的罚分，得到一个显著的分数是可能的。但是使用PAM10矩阵，氨基酸替换将得到较大的罚分，而且一般被认为是不可接受的。对两个近相关的lipocalin蛋白，它们有许多相同的氨基酸（如小鼠和大鼠RBP），对这类进化距离的蛋白，PAM10更合适。

A	7																			
R	−10	9																		
N	−7	−9	9																	
D	−6	−17	−1	8																
C	−10	−11	−17	−21	10															
Q	−7	−4	−7	−6	−20	9														
E	−5	−15	−5	0	−20	−1	8													
G	−4	−13	−6	−6	−13	−10	−7	7												
H	−11	−4	−2	−7	−10	−2	−9	−13	10											
I	−8	−8	−8	−11	−9	−11	−8	−17	−13	9										
L	−9	−12	−10	−19	−21	−8	−13	−14	−9	−4	7									
K	−10	−2	−4	−8	−20	−6	−7	−10	−10	−9	−11	7								
M	−8	−7	−15	−17	−20	−7	−10	−12	−17	−3	−2	−4	12							
F	−12	−12	−12	−21	−19	−19	−20	−12	−9	−5	−5	−20	−7	9						
P	−4	−7	−9	−12	−11	−6	−9	−10	−7	−12	−10	−10	−11	−13	8					
S	−3	−6	−2	−7	−6	−8	−7	−4	−9	−10	−12	−7	−8	−9	−4	7				
T	−3	−10	−5	−8	−11	−9	−9	−10	−11	−5	−10	−6	−7	−12	−7	−2	8			
W	−20	−5	−11	−21	−22	−19	−23	−21	−10	−20	−9	−18	−19	−7	−20	−8	−19	13		
Y	−11	−14	−7	−17	−7	−18	−11	−20	−6	−9	−10	−12	−17	−1	−20	−10	−9	−8	10	
V	−5	−11	−12	−11	−9	−10	−10	−9	−9	−1	−5	−13	−4	−12	−9	−10	−6	−22	−10	8
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

图 3-15 PAM10对数比值矩阵。低PAM值对比对近相关的蛋白质有用。把这个矩阵和PAM250矩阵比较（图3-14），并且注意在这个PAM10矩阵中等同匹配的分值比较大，不匹配的罚分也大

3.2.4　双序列比对中 PAM 矩阵的实际有用性

我们将通过使用 PAM40 和 PAM250 矩阵对两个蛋白进行双序列比对并且检查它们的结果，以阐述 PAM 矩阵的有用性。尝试一下基于 web 的 SIM 比对程序。ExPASy 提供了两个框，粘贴人 RBP4 和牛 β-乳球蛋白的序列（或者蛋白的 accession number）。注意你可以用下拉菜单（“comparison matrix”）来选择不同的 PAM 矩阵。开始我们尝试 PAM250，这个矩阵是合适的，因为两个蛋白仅是远程相关。在 81 个氨基酸的跨度上发现有 20 个相等残基的重叠（一致性为 24.7%）[图 3-16(a)]。接下来尝试 PAM40 矩阵。现在，最佳比对片段只有 10 个氨基酸长 [图 3-16(b)]。而且，这个短长度的比对在生物学上是无意义的。PAM40 矩阵对检测远相关蛋白质序列不合适。一个重要的概念是不同的打分矩阵对不同相关程度的蛋白质序列（或 DNA 序列）的敏感性不同。

> ExPASy 网站（http://www.expasy.ch）中包含了 SIM 比对程序（http://www.expasy.ch/tools/sim-prot.html）。

其他的双序列比对程序将在以后描述（见“3.7 网络资源”的表 3-5 和表 3-6）。

(a)

81 个重叠残基中 24.7% 的一致性：分值为 77.0：间隙频度为 3.7%

```
 hsrbp,  26 RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDV
btlact,  21 QTMKGLDIQKVAGTWYSLAMAASD-ISLLDAQSAPLRVYVEELKPTPEGDLEILLQKWEN
                  *     ****  *        * *       *               **  *

 hsrbp,  86 --CADMVGTFTDTEDPAKFKM
btlact,  80 GECAQKKIIAEKTKIPAVFKI
              **        *  ** **
```

(b)

10 个重叠残基中 60.0% 的一致性；分值为 23.0：间隙频度为 0

```
 hsrbp, 136 CRLLNLDGTC
btlact,   3 CLLLALALTC
            * ** *  **
```

图 3-16　RBP4 和 β-乳球蛋白双序列比对，使用（a）PAM250 矩阵和（b）PAM40 矩阵。检测远相关蛋白的相似区域，PAM250 矩阵较好。注意 PAM40 矩阵的比对结果不正确，它比对了 RBP4 的羧基端部分（自 136 位氨基酸开始）和 β-乳球蛋白的氨基端部分。“*”表示氨基酸相同

3.2.5　PAM 矩阵的重要替代者：BLOSUM 打分矩阵

除 PAM 矩阵外，另一个十分常用的打分矩阵是块替换矩阵（blocks substitution matrix）（BLOSUM）系列。Henikoff 和 Henikoff（1992，1996）使用 BLOCKS 数据库，它包括远相关蛋白序列的 500 组以上局部多序列比对（块）。这样，两位 Henikoff 的工作集中于远相关蛋白的保守区域（块）上。

BLOSUM62 矩阵把比对中所有氨基酸一致性≥62%以上的蛋白质合并成一条序列。如果比对的 RBP 直系同源物中的某块包含一些氨基酸一致性为 62%、80%和 90%的序列，那么它们将被合并成一条序列。BLOSUM62 矩阵的取代频率对一致性<62%的蛋白质序列的权重很大（这样，这个矩阵就可以用于评价一致性<62%的蛋白质）。BLOSUM62 矩阵是大多数 BLAST 算法的缺省矩阵（第 4 章），如图 3-17 所示。

Henikoff 和 Henikoff（1992）测试了一系列 BLOSUM 矩阵和 PAM 矩阵在 BLAST 搜索数据库时检测蛋白质的能力。他们发现 BLOSUM62 检测各种蛋白的效果比 BLOSUM60 和 BLOSUM70 稍

> BLOCKS 数据库位于 http://www.blocks.fhcrc.org。我们在第 8 章中会进一步讨论。

A	4																			
R	−1	5																		
N	−2	0	6																	
D	−2	−2	1	6																
C	0	−3	−3	−3	9															
Q	−1	1	0	0	−3	5														
E	−1	0	0	2	−4	2	5													
G	0	−2	0	−1	−3	−2	−2	6												
H	−2	0	1	−1	−3	0	0	−2	8											
I	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
L	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
K	−1	2	0	−1	−1	1	1	−2	−1	−3	−2	5								
M	−1	−2	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
F	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
P	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
S	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
W	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Y	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
V	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

图 3-17 Henikoff 和 Henikoff BLOSUM62 打分矩阵（1992）。该矩阵把所有比对中氨基酸一致性≥62%的蛋白质合并成一条序列。在检测远相关蛋白上，BLOSUM62 比其他 BLOSUM 矩阵和各种 PAM 矩阵效果要好。因此，它是大多数数据库搜索程序（如 BLAST，第 4 章）的缺省打分矩阵。该图授权使用

好，比 PAM 矩阵好得多。他们的矩阵对检测仅有微弱打分的比对特别有用。BLOSUM50 和 BLOSUM90 是 BLAST 搜索常用的另外矩阵（对一致性为 50%的两蛋白质的比对，试着使用 BLOSUM50 矩阵）。

PAM 矩阵由 10 乘以比值比以 10 为底的对数给出。BLOSUM 矩阵是由 2 乘以比值比以 2 为底的对数给出。因此，在相同尺度上 BLOSUM 分值不如 PAM 分值大。实际上，这个范围内的分值差异是不重要的，因为我们经常把原始分值转换成归一化的分值（normalized bit score）（第 4 章）。

PAM 矩阵和 BLOSUM 矩阵的相互关系可由图 3-18 大致描述。概括地说，BLOSUM 矩阵和 PAM 矩阵都在打分系统中使用对数比值。每种情况下，当需要执行一双序列比对（或者以一查询序列搜索数据库）时，应基于查询序列和匹配序列的一致程度，然后选择矩阵。PAM 矩阵是基于近相关蛋白家族数据的，并且假设高度相关蛋白的取代概率（如 PAM10）可以外推到远相关蛋白的概率（如 PAM250）。对比之下，BLOSUM 矩阵是基于

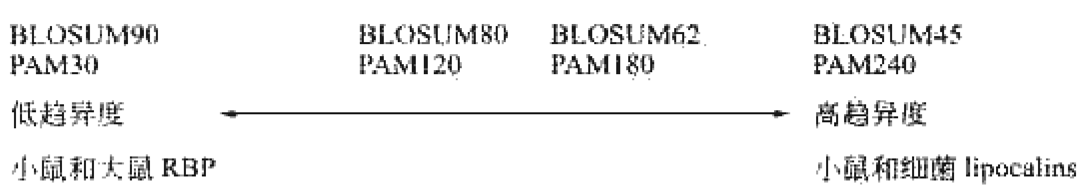


图 3-18 PAM 矩阵和 BLOSUM 矩阵概要。高值 BLOSUM 矩阵和低值 PAM 矩阵最适合于研究高度保守的蛋白（如小鼠和大鼠的 RBP）。低值 BLOSM 矩阵（如 BLOSUM45）或者高值 PAM 矩阵最适合于检测远相关蛋白。记住在 BLOSUM45 中，所有氨基酸一致性≥45%的家族成员被合在一起，它允许矩阵集中于一致性<45%的蛋白

实际观测到的远相关蛋白比对。

3.2.6 双序列比对和检测限度

当比较两蛋白序列时，在它们的关系不能辨认之前允许多少突变发生？当比较人和鲑鱼 RBP 时，很容易看到它们的近相关性（见图 3-8）。但是，当我们比较 RBP 和牛 β-乳球蛋白时，关系变得比较不明显（图 3-5）。直观上存在某一点，在这一点上两个同源蛋白的差异程度较大使得它们的比对结果没有显著性。

双序列比对的有用检测限度能通过比较两序列的一致百分比（或者趋异百分比）和它们的进化距离来阐述。假设两条蛋白序列各为 100 个氨基酸的长度，一条序列固定，另一条序列引入不同数目的突变，那么两条序列分化的图形趋势呈一种负指数形式（图 3-19）（Doolittle，1987；Dayhoff，1978）。如果两条序列有 100%的氨基酸一致性，则它们每 100 个残基中没有发生改变。如果它们有 50%的氨基酸一致性，则它们每 100 个残基中平均发生 80 个变化。在这个 50%的一致性的情况下，或许有人认为每 100 个残基中平均有 50 个改变。但是任何一个位置都可能发生多次改变。因此，一致性百分比不是蛋白质序列中发生突变数目的确切指标。当一个蛋白每 100 个氨基酸发生 250 次变化（hit）时，它可能会和原蛋白质有 20%的一致性，这仍能被认为是显著性相关的。如果蛋白质每 100 个氨基酸发生 360 次变化，那么它已经到了两个蛋白有大约 15%氨基酸一致程度的点上，因此它们已不再可能被认为是显著相关的。

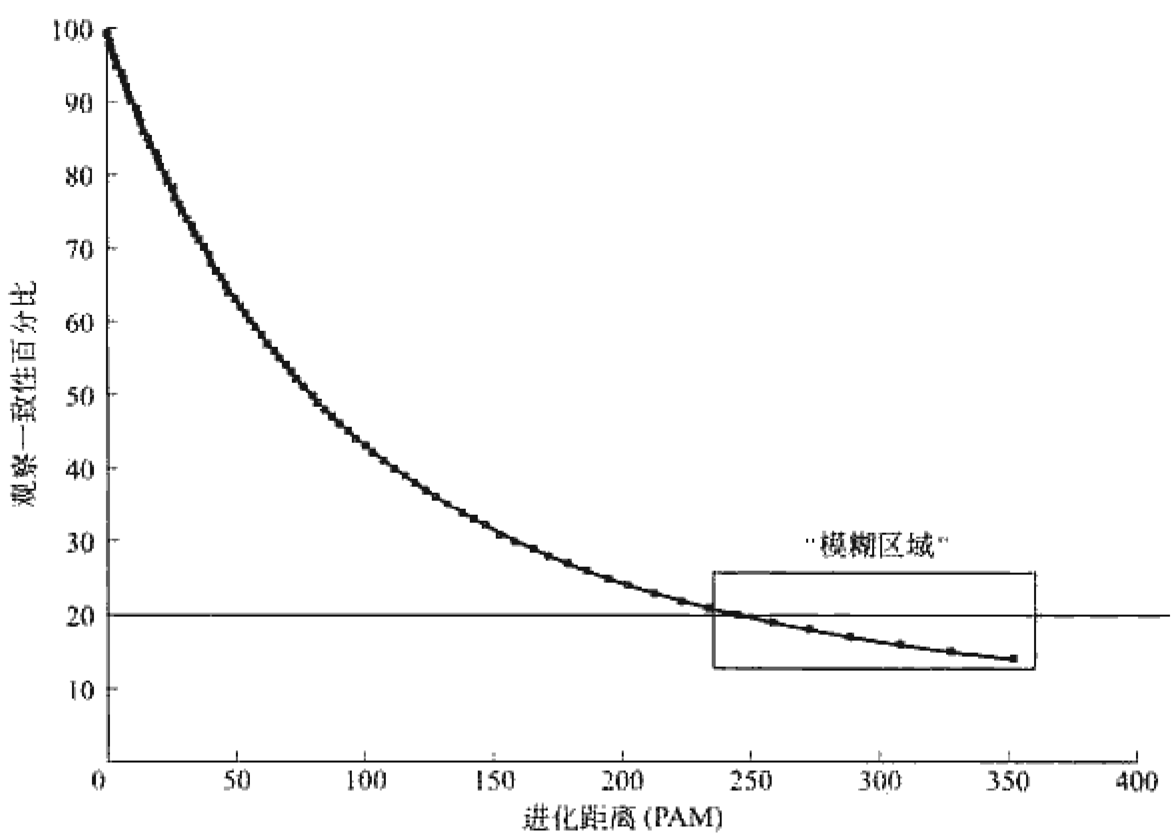


图 3-19 两随机趋异蛋白质序列以负指数方式变化。该图显示了两序列每百个残基中相同残基的数目（y 轴）和它们发生突变的数目（以 PAM 为单位的进化距离）的关系。模糊区域（Doolittle，1987）的进化距离对应两蛋白 20%的氨基酸一致性。具有这种尺度上的氨基酸一致性的蛋白可能是同源的，但是同源性难以检测。该图使用来自 Dayhoff（1978）（见表 3-4）的数据（由 Matlab 所画）

PAM250 矩阵假设每 100 个氨基酸发生 250 次点突变（图 3-19），这对应于“模糊区域”。在这个趋异水平上，要评价两蛋白的同源性通常比较困难。其他技术，包括多序列比对（第 10 章）和结构预测（第 9 章），对评价这种情况下的同源性可能有用。PAM 矩阵存在从 PAM1 到 PAM250 的情况，每 100 个残基中观察到的氨基酸的不同数目与每个 PAM 矩阵相关（表 3-4 和图 3-19）。

一次变化（hit）表示由突变导致的一次氨基酸的改变。我们将在第 11 章讨论突变（包括同一个核苷酸位置的多次改变）（见图 11-11）。我们将在第 18 章讨论与人类疾病相关的突变。

图3-19中的曲线在大约15%的氨基酸一致性以下时趋于一渐近线。如果比较蛋白时不允许间隙，那么这条渐近线将趋于5%（或者氨基酸的平均背景频率）。

表3-4 两比对蛋白质序列每百个残基中观察到的有差异氨基酸的数目和进化差距的关系①

每百个残基中观察到的差异数目/个	进化距离/PAM	每百个残基中观察到的差异数目/个	进化距离/PAM
1	1.0	45	67
5	5.1	50	80
10	10.7	55	94
15	16.6	60	112
20	23.1	65	133
25	30.2	70	159
30	38.0	75	195
35	47	80	246
40	56		

① 发生变化的数目以PAM为单位，源自Dayhoff（1978），授权使用。

3.3 比对算法：全局和局部

至此，我们的讨论集中于给两蛋白质比对打分的矩阵。这包括产生相等匹配、不匹配和间隙的打分值。我们也需要一个合适的算法来执行比对。当两个蛋白质比对在一起时，存在数量巨大的比对可能。

两条长为 n 的序列，存在大约 $2^{2n}/\sqrt{\Pi n}$ 种可能的全局比对（Durbin 等，2000）（在这个方程中，Π是乘积符号）。对两条长为1000的序列，存在 10^{600} 种比对可能。对两条长为200氨基酸残基的蛋白质（如lipocalin），可能比对的数目为 6×10^{58}（6×10^{58} 为一指数形式，它也能表示成6E58或者6exp58）。

主要有两类比对：全局和局部。我们接下来将讨论这些算法。全局比对（如Needleman-Wunsch方法）包含每个蛋白或DNA的全部序列。我们已经在图3-5中看到过RBP4和β-乳球蛋白的全局比对。局部比对（如Smith-Waterman方法），集中于两序列间最大相似的区域。许多情况下，我们选择局部比对，因为仅有两蛋白的一部分需要比对（我们将在第8章研究蛋白质的模块化特点）。许多数据库搜索算法（如BLAST，第4章）使用局部比对算法。

每种方法都能保证找到一种或多种蛋白质或DNA序列比对的优化解。我们将描述两种快速搜索算法，即BLAST和FASTA。BLAST是一种局部比对的简化形式，较受欢迎，因为算法较快而且容易得到结果。

3.3.1 全局序列比对：Needleman-Wunsch算法

首先用于蛋白质序列比对的重要算法之一是由Needleman和Wunsch（1970）提出的。这个算法较为重要，因为它能产生蛋白质和DNA序列的最优化比对，甚至允许引入间隙。该算法结果是最优化的，但没有穷举所有可能比对。一个穷尽的双序列比对算法需要太多的计算时间以至于不可能得到结果。

我们分三步介绍全局序列比对的Needleman-Wunsch算法：①设置一个矩阵；②给矩阵打分；③确定最优比对。下面我们描述这个过程，使用的矩阵是30多年前原始文献中所描述的（Needleman和Wunsch，1970）。

步骤一：设置一个矩阵。首先，我们在一个二维矩阵中比较两条序列（图 3-20 及其以下的图）。第一条序列长 m，沿 x 轴水平排列，这样，各氨基酸残基对应于每列。第二条序列长 n，沿 y 轴垂直排列，其氨基酸残基对应于每行。

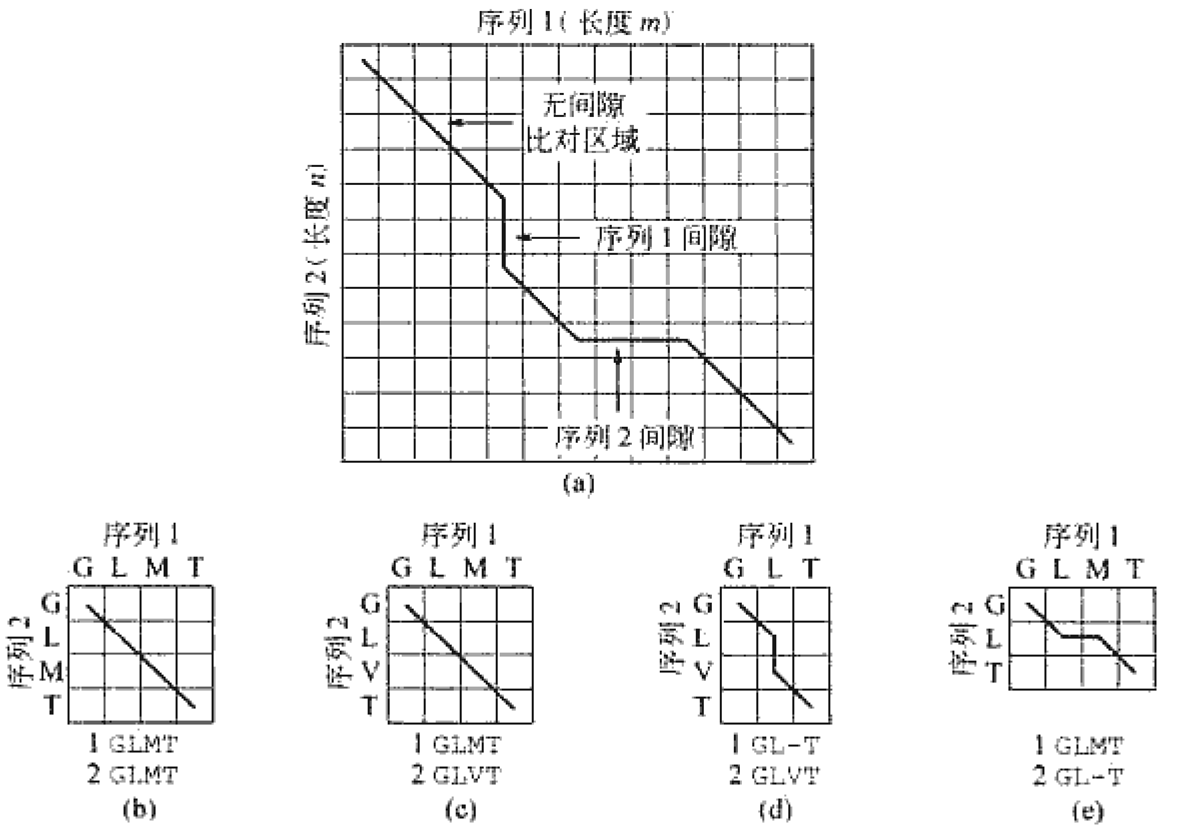


图 3-20　使用 Needleman-Wunsch（1970）动态规划算法为两条氨基酸序列执行全局双序列比对。(a) 两条序列可以指定到一穿过矩阵的对角路径上，如果有必要，这条路径可以有水平线和垂直线，表示比对中引入了间隙；(b) 两条等同序列的路径正好是矩阵的对角线；(c) 如果存在不匹配（或多个不匹配），虽然路径仍是对角的，但打分系统会给这种不匹配提供罚分；(d) 第一条序列或 (e) 第二条序列路径包含了垂直线或水平线

我们在下面会描述寻找一条穿过矩阵的对角路径的规则。这条路径就代表了两序列的比对。两条等同序列的最佳比对可由从左上角到右下角的对角线简单表示 [图 3-20(a)，图 3-20(b)]。两条序列的任何不匹配也会表示在这条对角路径中 [图 3-20(c)]。但是，所给的分值必须根据打分系统来调整。在图 3-20(c) 的例子中，残基 V 和 M 不匹配，所给的分值比 M 和 M 完全匹配低 [图 3-20(b)]。

间隙在这个矩阵中表示为水平或垂直路径，如图 3-20(a)，图 3-20(d)，图 3-20(c) 所示。水平序列的间隙为一垂直线 [图 3-20(a)，图 3-20(d)]，垂直序列的间隙由一水平线表示 [图 3-20(a)，图 3-20(e)]。间隙可以有任何长度，可以是中间的或端部的。

我们下一步是设置矩阵用于比较两条序列，其方法同 Needleman 和 Wunsch（1970）原始论文描述的一样（图 3-21）。我们将详细研究这个例子，以明白怎样全局地比对两条序列。第一条序列沿 x 轴水平排列，第二条序列沿 y 轴垂直排列。

Needleman-Wunsch 方法是一种动态规划算法。称作"动态"是因为比对是基于逐个残基搜索的最优比对。"规划"指使用一系列规则来确定这个比对。

步骤二：给矩阵打分。这个算法的目标是确定一个最佳比对。如下所述，我们需要设置两个矩阵：氨基酸相等矩阵和打分矩阵。在打分矩阵，我们采用一个简单的打分系统，氨基酸相同得分为 +1，每一次不匹配得分为 −1，出现一个间隙位置得分为 −2。

这个算法有时也被称作 Needleman-Wunsch-Sellers 算法。Sellers（1974）提出了一个相关的比对算法（该算法集中于最小化差异性，而不是最大化相似性）。Smith 等（1981）指出 Needleman-Wunsch-Sellers 方法在数学上是等价的。

我们首先建立氨基酸相等矩阵。如果两个蛋白的氨基酸残基相同，那么就在矩阵对应的位置放上 +1 [图 3-21(a)]。在图 3-20(b)，如果两序列等同，矩阵内 +1 值占据了从左上到

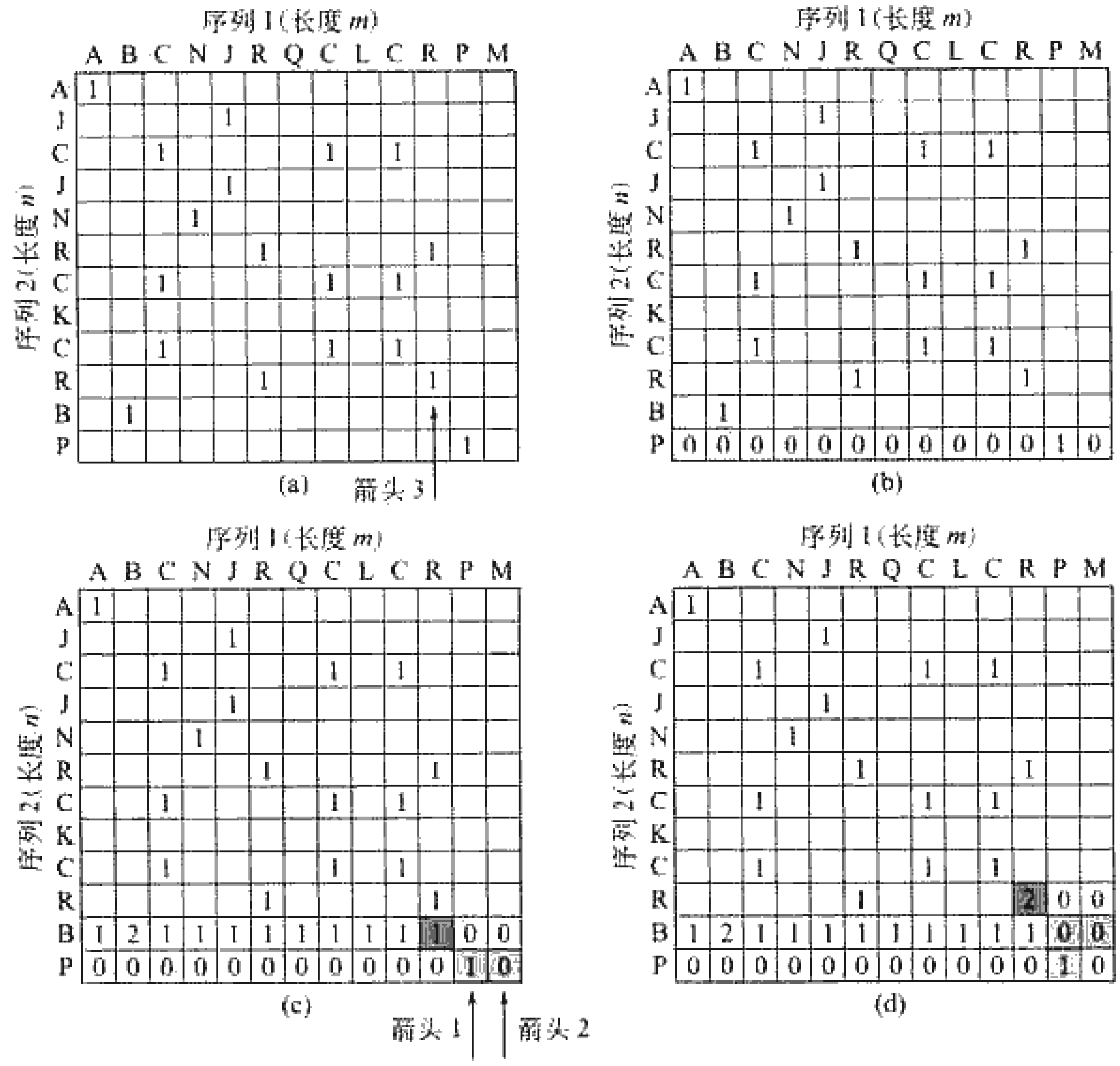


图 3-21 使用 Needleman-Wunsch (1970) 动态规划算法对两氨基酸序列的全局双序列比对这个例子中的序列来自于原始文献。(a) 我们先在氨基酸相等矩阵中等同氨基酸的对应位置赋值为+1，其他单元格赋值为0（没显示）。从每个蛋白的羧基端（矩阵的右下角）开始，我们逐行计算矩阵值。(b) 底部一行，我们赋值为0，除了 $P\times P$ 单元格赋值为+1。遵循的规则为，每个单元格的值为+1［如果在（a）中存在］加上该单元格的对角右下单元格及其下边列、右边行单元格值的最大值。在该图和图3-22，这表示从灰色阴影单元格中选取最大值。(c) 底部上面一列 $B\times R$ 单元格（深灰色阴影）的分值为+1，因为这个值是对角右下单元格（$P\times P$ 单元格，灰色阴影）和它的右边行单元格（灰色阴影）值的最大值。该单元格的值+1被加到第二行左边的所有单元格上。注意在（c）图中第二行 $B\times B$ 单元格的分值为+2（+1来自 $B\times B$ 的等同匹配，另一+1来自下面一行 $P\times P$ 的匹配）。(d) 至于 $R\times R$ 单元格（红色阴影），它的值为+1（来自相等矩阵）加上其右下三个值（灰色阴影）中的最大值（加+1）。此图已被授权使用

右下的对角线。

接下来，我们设置打分矩阵［图 3-21(b)～图 3-21(d)］。我们的目标是找到最佳比对，确定最大化这个分值的路径。这通常需要找到的路径在其尽可能多的位置上有等同氨基酸，而间隙尽可能地少。在每个位置（i, j，即矩阵的每个单元）有4种可能：

① 两残基完全匹配（即相等）；

② 它们不匹配；

③ 第一条序列引入一个间隙；

④ 第二条序列引入一个间隙。

Needleman-Wunsch 算法为比对序列的每个位置上对应于上述出现的每种可能提供了一个分值。算法也提供了一套规则来描述我们在矩阵中如何移动。

分值的赋值从矩阵的右下部开始，对应于蛋白质的羧基端。我们给每行赋值，每次一个单元格，从最下面一行右边开始行至顶部。最下面一行，我们赋值 0，除了 $P \times P$ 的单元格赋值为 +1 [图 3-21(b)]。我们给 $P \times P$ 单元格赋值 +1 是因为在氨基酸相等矩阵中如此 [图 3-21(a)]。

简单打分系统和氨基酸取代（分值—1）相比，插入和删除（分值—2）较少，因此它对插入和删除惩罚较重。

我们如何在矩阵中移动、并定义一条路径，使它对应于序列比对？定义一矩阵，i 代表行，j 代表列。一个单元格定义为位置 (i,j)。对于任何一个我们想确定其分值的单元格 [如在图 3-21(c) 中带红色阴影的单元格] 都有以下几条规则。

请注意在线性代数中，相等矩阵（即单位矩阵，identity matrix）是一类特殊的矩阵，左上至右下对角线上的值为 1。对于序列比对，氨基酸相等矩阵是表示两条序列氨基酸相同的位置的简单矩阵，如图 3-21（a）所示。

① 首先，i 和 j 都必须增长。这意味着为了计算深灰色单元格的值，我们必须到矩阵内该单元格的下面和右面去找。这是合理的，因为我们不能破坏序列内氨基酸（或核苷酸）的线性排列。

② 如果 i 和 j 其中之一增长 1，则另一个可以增长 1 以上。在这个例子中，i 只能增长 1。单元格的 j 值可以增长 1 [图 3-21(c)，灰色，箭头 1 所示]，也可以增长 2（灰色单元格，箭头 2 所示）。

我们分两步给每个单元格打分。我们检查氨基酸相等矩阵 [图 3-21(c)] 并从中取值 [该例子中，(c) 图为深灰色单元格，其值为 0]。然后我们使用上述两条移动规则，加上得到的最大值。在本例中，该单元格值为 1 [图 3-21(c)，箭头 1 所示]。因此，给图 3-21(c) 的深灰色单元格赋值为 1。

如果我们检查矩阵中已赋值的其他单元格，移动规则可以变得更简单。对于图 3-21(d) 的深灰色单元格，其值为 +1 [来自氨基酸相等矩阵，图 3-21(a)，箭头 3 所示] 加上移动规则得到的最大值 +1。这样，我们在该深灰色单元格中赋值 2。下一个单元格 [图 3-22(a)]，其值为从氨基酸相等矩阵得到的值 0 加上阴影单元格的最大值 1，即总分值为 1。

作为例子，图 3-22(b) 中深灰色阴影单元格的分值为 +5。这个分值为 +1（来自氨基酸相等矩阵）加上 +4（通过我们的移动规则得到的最大值）。+4 值位于 $(i+1, j+2)$ 单元格。对这个“跳跃”（比对角线 $i+1, j+1$ 上 +3 的分值高）的解释是，在序列 1 [残基 RQC，图 3-22(c)，图 3-22(d)，图 3-22(e)] 和序列 2（残基 R-C）的比对中引入间隙得到一个高分值，因而是一最佳比对。

概而言之，每个位置 (i,j) 的分值为该单元格的分值加上下列三个值中的最大值。

① 对角线下方（位置 $i+1, j+1$）的分值。不用引入间隙。

② 从位置 $(i+1, j+2)$ 到其行尾的最大值。这对应于列中引入一间隙。间隙可以在一个或多个位置。

③ 从位置 $(i+2, j)$ 到其列末的最大值。对应于在行中加入一间隙。

步骤三：确定最佳比对。矩阵赋值完毕后，可以通过一回溯方法确定比对。在此处，我们从矩阵的左上（蛋白质的氨基端）开始，即从最大值 +8 开始 [图 3-22(c)]。这对应两序列中残基 A 和 A 比对 [图 3-22(a)～图 3-22(e)]。我们往下往右找最大值。接下来的比对是 J 和 B（不匹配，分值为 +7）。下一步是两相同的 C 残基。每一步，我们都循着对角线及其右下方寻找最大值。在下一个位置，可能是对角线上的 J 和 N 比对，其分值为 +5，但是选

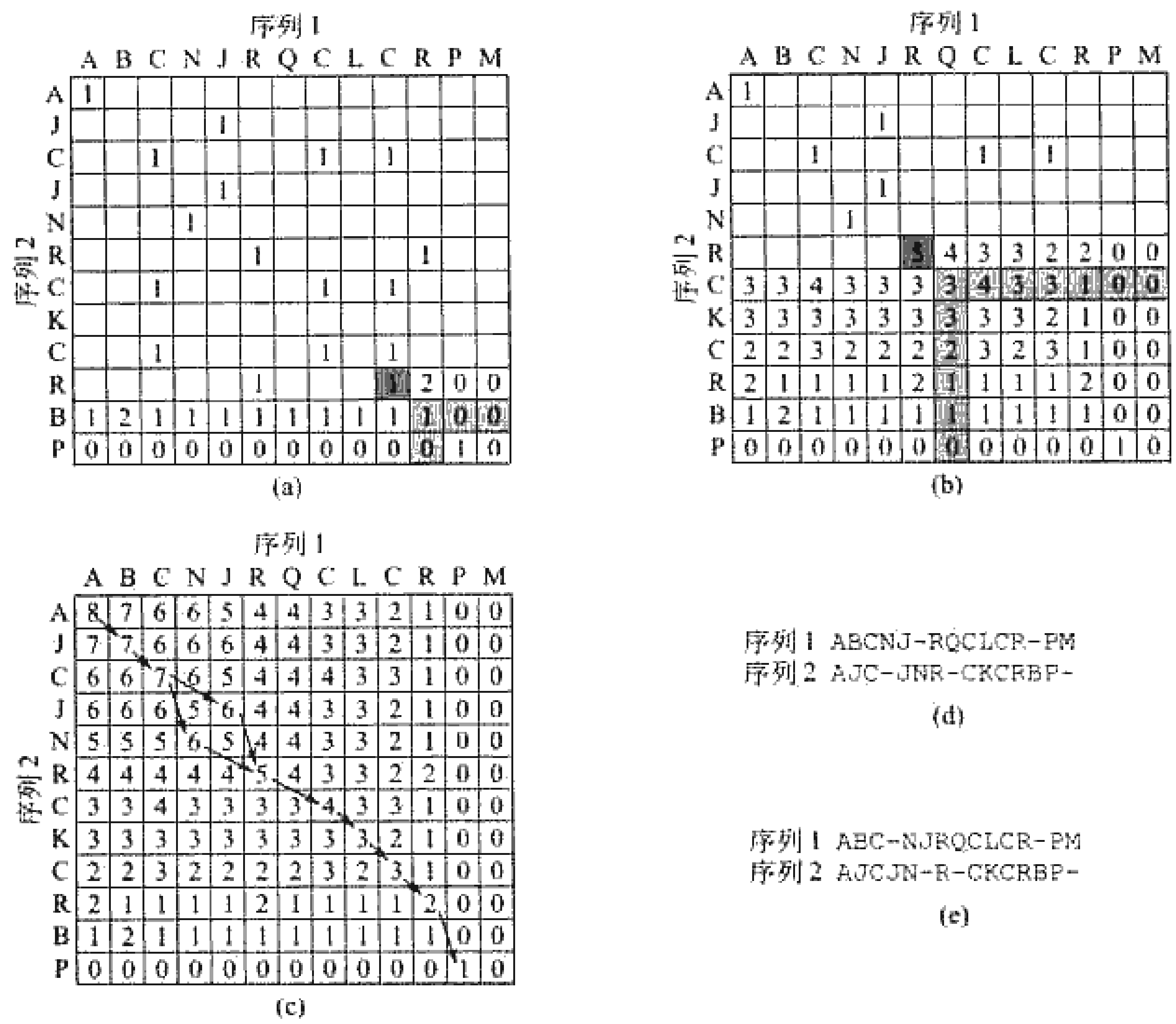


图 3-22 使用动态规划算法对两氨基酸序列全局双序列比对：给矩阵打分和执行回溯过程以得到比对结果。矩阵打分（图 3-21）逐行向上进行。(a) 在深灰色阴影单元格，分值是 0（没有相等氨基酸分值为 +1）加上灰色单元格的最大值（+1）。(b) 在 $R\times R$ 单元格（深灰色阴影），分值为 +1（相同氨基酸）加上方括号中三个分值中的最大值 [$(i+1, j+1)$（分值为 +3），$(i+2, j+1)$ 至 i 行行尾（分值为 +4），$(i+1, j+2)$ 至该列的底部]。因此最佳分值是 +1 + (+4) = +5。(c) 通过这个方法给矩阵赋值后，使用回溯过程来确定最佳的比对路径 (s)。从每个蛋白的氨基端开始（分值为 +8），然后沿着分值穿过矩阵。此处，开始的三个残基的比对没有间隙（第一条序列 ABC，第二条序列 AJC）。有一处出现两条同样最优的路径，每条的分值都为 +6，都在其中一条序列中引入了一个间隙。上面一条路径显示于 (d)，下面一条路径显示于 (e)。因为是全局比对，所以整条序列都参与了比对，而不仅是内部的一个片段。此图已被授权使用

择右边或下边单元格能得到更高分值（+6）[图 3-22(d)，图 3-22(e)]。偏离对角线意味着要在一条序列中加入间隙，这对应一些罚分。可能有多于一个的最佳比对（如同这个例子显示）存在两个答案 [图 3-22(d)，图 3-22(e)] 有同样高的分值。这里的打分系统仅包括匹配的分值（+1）。但是，如引入复杂的打分系统（如 BLOSUM62 或 PAM250），则不大可能存在多个最佳比对。

有一系列程序实现了全局比对算法（见本章结尾的“Web 资源”）。例子之一就是 GCG（框 3-6）的 GAP 程序。输入两条序列，然后选择间隙罚分和延长罚分的值（图 3-23）。全局比对的结果包括两条蛋白的一致性百分比和相似百分比、比对长度和引入间隙的数目（图 3-24）。

Needleman-Wunsch 算法是动态规划的一个例子（Sedgewick，1988）。这表示最佳路径

（即最佳比对）是通过逐步延长最佳子路径得到的，即比对的每一步都选择获得最高分值的残基对作为比对方法，总目标是沿着矩阵对角线找到一条获得最大分值的路径。这条路径确定了最佳比对。

框 3-6 Genetics Computer Group

Genetics Computer Group（GCG）也称作“Wisconsin package”（http://www.accelrys.com）（见附录）。GCG 集合了分析 DNA、RNA 和蛋白质序列的 130 种算法。如今全世界有 4 万多 GCG 用户。这是一个商业软件包，有 Unix 下和基于 web 的版本。

使用 GCG，用户首先必须把序列放入主要的序列编辑器“seqed”。蛋白质和 DNA 序列能手工输入或者通过“fetch”命令获得。（对于所有 GCG 命令，通过“genhelp”工具都能得到详尽的帮助菜单。）

一旦相关序列在 GCG 工作目录中准备好之后，可以使用“GAP”工具（使用 Needleman-Wunsch 算法的全局序列比对）和“Bestfit”工具（使用 Smith-Waterman 算法的局部序列比对）来执行双序列比对。我们先用 GAP 进行全局比对（见图 3-29）。注意程序允许用户选择所比较氨基酸的范围和间隙权重和长度的罚分。输出（见图 3-30 以后）包括品质分值（quality score），它基于所使用的打分矩阵，一个高的品质分值对应于更显著的比对。其他数据包括氨基酸一致性百分比，间隙的数目和比对自身。对于全局比对，程序会显示两蛋白的全部序列，除非（这个例子）蛋白质的羧基端不存在匹配。

局部双序列比对的方法类似，使用“Bestfit”程序。我们在命令行下输入“bestfit hsrbp.pep btlacto.pep ran=100”，它指示 GCG 运行 bestfit 程序，比较我们的两个蛋白。命令中“ran=100”意识是第二条序列随机排列并和第一条序列比较 100 次。我们注意输出的几个特点（图 3-26 以后）。比对是局部的，所以比对仅显示了双序列比对中较好匹配的部分。比对区域比全局比对短，但是匹配区域的一致性百分比要大。全局和局部比对的结果经常会差异很大。对双序列比对和多序列比对都是如此（第 10 章）。

```
> gap

Gap uses the algorithm of Needleman and Wunsch to find the alignment of
two complete sequences that maximizes the number of matches and minimizes
the number of gaps.

 GAP of what sequence 1 ?  hsrbp.pep

                  Begin (* 1 *) ?
                End (*   199 *) ?

 to what sequence 2 (* hsrbp.pep *) ?  btlacto.pep

                  Begin (* 1 *) ?
                End (*   178 *) ?

 What is the gap creation penalty (* 8 *) ?

 What is the gap extension penalty (* 2 *) ?

 What should I call the paired output display file (* hsrbp.pair *) ?

 Aligning ........-.
 Aligning ........-.

         Gaps:      0
      Quality:     37
 Quality Ratio: 0.208
  % Similarity: 31.902
        Length:   214
```

图 3-23 使用 GCG 的 GAP 程序执行全局双序列比对。在这个例子中，我们比较人 RBP4（hsrbp.pep）和牛 β-乳球蛋白（btlacto.pep）。序列从 NCBI Entrez 的蛋白质站点拷贝到 GCG 序列编辑器（“seqed”），然后以 .pep 作为后缀命名，以表示它们是蛋白质而不是核酸。一旦输入 GCG 后，序列能以数十种不同的算法来分析（包括 GAP）。这里，蛋白质比较的区域和间隙引入延长的罚分都可以由用户选择


```
     Gap Weight:        8      Average Match:  2.912
  Length Weight:        2   Average Mismatch: -2.003

        Quality:       37             Length:    214
          Ratio:    0.208               Gaps:      8
Percent Similarity: 31.902   Percent Identity: 26.300

      Match display thresholds for the alignment(s):
                  | = IDENTITY
                  : =    2
                  . =    1

hsrbp.pep x btlacto.pep    July 16, 2001 14:45  ..

    1 MKWUWALLLLAAWAAAERDCRUSSFRUKENFDKARFSGTWYAMAKKDPEG 50

    1 ...MKCLLLALALTCGAQALIUT..QTMKGLDIQKUAGTWYSLAMAASD. 44

   51 LFLQDNIUAEFSUDETGQMSATAKGAUR.LLNNWD..UCADMUGTFTDTE 97

   45 ISLLDAQSAPLAU.YUEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93

   98 DPAKFKMKYWGUASFLQKGNDDHWIUDTDYDTYAU..........QYSC 136

   94 IPAUFKIDALNENKUL........ULDTDYKKYLLFCMENSAEPEQSLAC 135

  137 ALLNLDGTCADSYSFUFSRDPNGLPPEAQKIUAQRQ.EELCLARQYRLIU 185

  136 QCLUATPEUDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI....... 178
```

图 3-24 使用 GAP 对人 RBP4 和牛 β-乳球蛋白进行全局双序列比对的输出。输出项目包括一个分值的描述（quality=37）、间隙的数目和一致性百分比。在显示的比对中，第一条序列（RBP4）位于上方

3.3.2 局部序列比对：Smith-Waterman 算法

局部序列比对 Smith-Waterman 算法（1981）是两条蛋白质和 DNA 序列部分比对的最严格的算法。局部序列比对在许多情况下相当有用（例如数据库搜索），在那些情况下我们需要比对蛋白质的结构域（而不是整条序列）。局部序列比对算法和全局比对相似，两条蛋白质被排列在一个矩阵中，然后沿对角线搜索最佳路径。但是，从中间某个位置开始的比对不存在罚分情况，比对不需要延至序列的两端。

对于 Smith-Waterman 算法，所构造的矩阵在顶部多 1 行和在左边多 1 列。这样，对长度为 m 和 n 的序列，矩阵有 $m+1$ 和 $n+1$ 维。Smith-Waterman 算法决定矩阵每个位置分值的规则和 Needlemna-Wunsch 算法中的稍微不同。每个单元格的分值由前面对角线的分值和引入一间隙得到的分值的最大值决定。但是，分值不能为负（这是 Smith-Waterman 算法引入的一条规则），如果所有其他打分选择都产生了负值，那么在这个单元格上赋值为零。这样，分值 $S(i,j)$ 由下面四个值的最大值给出（图 3-25）。

① 单元格 $i-1$，$j-1$ 的分值；即对角线左上的分值。这个分值，加上单元格 $S[i, j]$ 的分值（包括匹配和不匹配）。

② $S(i, j-1)$（即左边单元格）减去间隙罚分。

③ $S(i-1, j)$（即上面单元格）减去间隙罚分。

④ 零。这个条件保证矩阵中没有负值。与全局比对不同，那里常出现负分，这是因为间隙罚分和不匹配的罚分（注意本章中的对数比值矩阵）。

图 3-25 显示了使用局部序列比对来比对两条核酸序列，它们取自 Smith 和 Waterman 的研究（1981）。矩阵的顶部行和最左边列赋值为零。最大比对可以从矩阵中的任意位置开始和结束（两蛋白质序列的线性顺序不能被破坏）。这个过程是找出矩阵中的

序列 1

		C	A	G	C	C	U	C	G	C	U	U	A	G
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A	0.0	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
U	0.0	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
G	0.0	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
C	0.0	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3
C	0.0	1.0	0.7	0.0	1.0	3.0	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0
A	0.0	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.3	0.7	1.0	1.3	1.3	0.0
U	0.0	0.0	0.7	1.7	0.3	1.3	2.7	2.3	1.0	0.7	1.7	2.0	1.0	1.0
U	0.0	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0	0.7	1.7	2.7	1.7	1.0
G	0.0	0.0	0.0	1.3	0.0	1.0	1.0	2.0	3.3	2.0	1.7	1.3	2.3	2.7
A	0.0	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3	2.0
C	0.0	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0	2.0
G	0.0	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0
G	0.0	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0

序列 2

(a)

序列 1 GCC-UCG
序列 2 GCCAUUG

(b)

序列 1 CAGCC-UCGCUUAG
序列 2 AAUGCCAUUGACGG

(c)

图 3-25 Smith 和 Waterman 局部序列比对（1981）。(a) 这个例子中，矩阵来自两条 RNA 序列（CAGCCUCGCUUAG 和 AAUGCCAUUGACGG）。这个不是相等矩阵［如图 3-21(a) 所示］，相等核苷酸的位置为灰色阴影。这里的打分系统是，匹配＋1，不匹配－1/3，间隙罚分为不匹配和匹配的差值（一个间隙－1.3）。矩阵打分的规则由文中给出。矩阵中的最大值 3.3 对应于最佳局部比对的开始，比对的残基（灰色阴影）向上向左延伸，直到碰到一个零分值。(b) 由矩阵得到的局部序列比对。注意这个比对包括匹配、一不匹配和一间隙。(c) 显示了两条序列的全局比对，以和局部比对作比较。注意它包含了两条序列的全部。本图已被授权使用

最大值［图 3-25(a) 中的值为 3.3］。这代表了比对的结尾处（比对部分的羧基端）。其位置不必在右下角，这在全局比对中也是如此。回溯过程是从最大值的位置开始，沿对角线向上向左直到碰到一个零分值的单元格。这个过程确定了比对起始，它也不必位于矩阵的顶部。

Smith-Waterman 算法需要的一个条件就是随机匹配的期望分值为负。这个条件保证了两条不相关的长序列不能得到高分值。否则，这样的序列产生的虚假比对得到的分值比两蛋白短区域内的真实比对的分值要高。

图 3-26 显示了使用 Smith-Waterman 算法对 RBP4（hsrbp.pep）和牛 β-乳球蛋白局部序列比对的例子。把它和图 3-24 全局比对的结果比较，注意局部比对的比对区域较短，但是一致性百分比和相似百分比较大。

局部序列比对使用的是 GCG 的 Bestfit 程序（http://www.accelrys.com）。

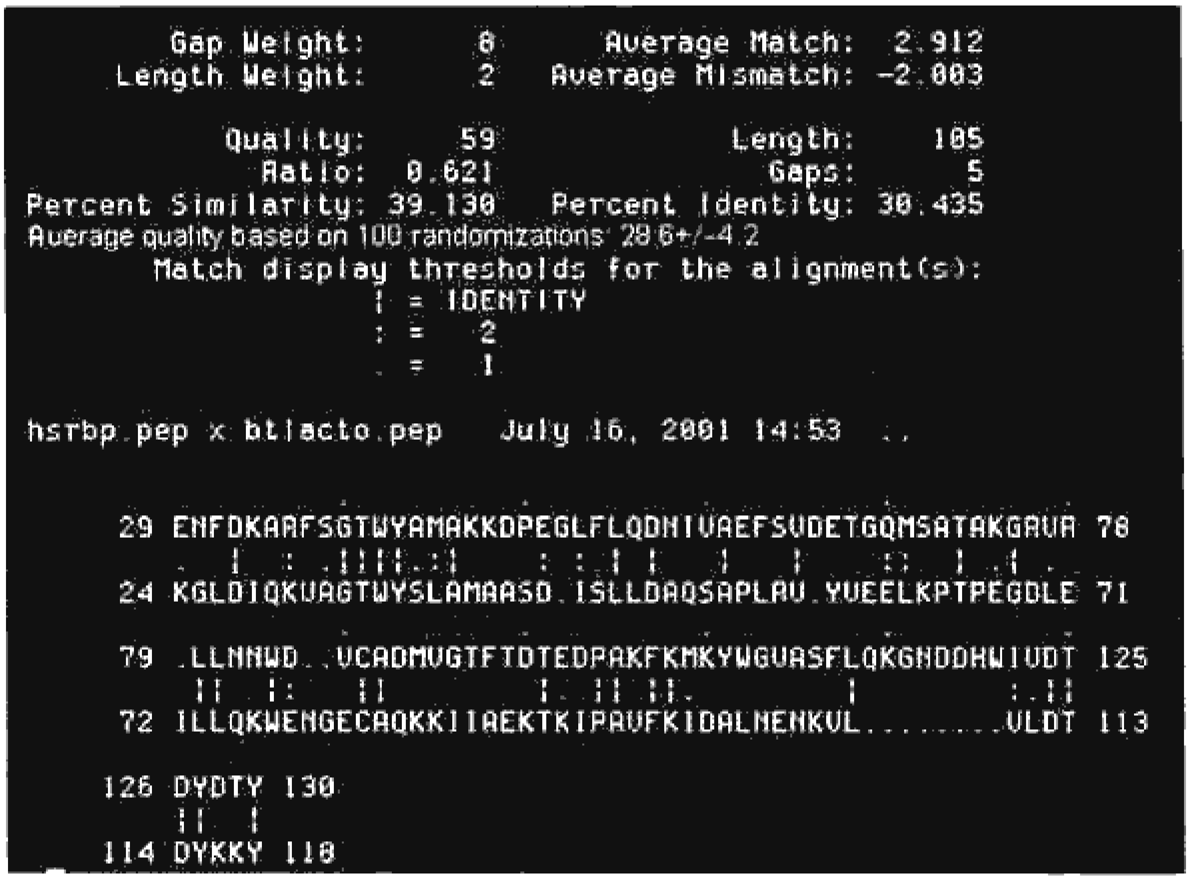


图 3-26 使用GCG的Bestfit程序局部比对，显示了双序列比对的有限区域。把这个比对和图3-24的比较，那里使用GAP算法的全局比对使得输出包含了两蛋白的整个序列。这里，比对的长度短，但是一致性百分比高。注意命令行中的"-ran=100"用来为100个随机排列的β-乳球蛋白序列产生品质分值（正或负标准偏差）。参见图3-30

3.3.3 Smith-Waterman算法的快速、启发式版本：FASTA和BLAST

Smith-Waterman算法能保证找到两条序列的最佳比对，但是它的速度慢。对于双序列比对，速度不是一个问题。但是当双序列比对应用于一条序列（查询序列）和整个数据库比较时，算法的速度成为一个严重的问题，并且可能会千差万别。

许多算法有一个参数 N，指它所处理的数据数量（见Sedgewick，1988）。这个参数很大程度上影响算法执行一个任务所需要的时间。如果任务所需时间和 N 成比例，那么 N 加倍时所需时间也成倍。如果任务所需时间是 N 的二次方（N^2），那么 $N=1000$ 时，执行时间就是百万量级。对于Needleman-Wunsch和Smith-Waterman算法，比对两条序列计算机所需的空间和时间都至少和两条序列长度的乘积 $m\times n$ 成比例。搜索一个大小为 N 的数据库，这个值就是 $m\times N$。

> Gotoh（1982）与Myers和Miller（1988）改进的算法需要 $O(mn)$ 时间和 $O(n)$ 空间。算法不把整个矩阵载入内存，并且忽略了小于某个阈值下的分值，使算法专注于搜索过程中能得到的最大分值的部分。

另外一种描述方法是O-表示法（O-notation，称作"big-Oh notation"），它近似一个算法执行时间的上限。Needleman-Wunsch算法需要 $O(mn)$ 步骤，Smith-Waterman算法则需 $O(m^2n)$ 步骤。后来，Gotoh（1982）和Myers和Miller（1988）改进了这个算法，使它所需时间和空间变小。

> FASTA表示FAST-All，指它对所有序列（即蛋白质和核酸）快速比对的能力。

两个流行的局部序列比对算法可以作为替代Smith-Waterman的快速版本：FASTA（Pearson和Lipman，1998）和BLAST（Basic Local Alignment Search Tool）（Altschul等，1990）。它们需要的时间较少。节省时间是因FASTA和BLAST在执行严格的比对之前先扫描数据库以发现可能的匹配序列。这是一种启发式算法（框3-2），牺牲了灵敏度来换取速度。与Smith-Waterman算法不同，它们不能保证找到最佳比对。

由Pearson和Lipman（1988）提出的FASTA搜索算法分如下4步。

① 生成一个查询表，包括数据库中短的氨基酸和核苷酸片段。短片段的长度由参数 *ktup* 决定。如果蛋白质搜索 *ktup*=3，查询序列就以 3 个氨基酸为块到查询表中查找可能的 3 个氨基酸匹配。FASTA 程序对一个给定的 *ktup* 值，产生 10 个最高分值片段。

② 这 10 个匹配区域重新打分，并允许有保守性替换，打分过程中使用打分矩阵（如 PAM250）。

③ 属于同一蛋白质的高分值区域连接在一起。

④ FASTA 然后在高分值序列上执行全局（Needleman-Wunsch）或者局部（Smith-Waterman）比对，这样可以优化查询序列和数据库匹配序列的比对。动态规划算法应用于数据库时只得到有限的使用，允许 FASTA 能非常快速地返回结果，因为它只评估了一部分潜在的比对。

> 参数 *ktup* 意指多元，例如二元、三元或四元（$k=2, k=3, k=4$）。*ktup* 的值通常是 3～6（核苷酸序列）和 1～2（氨基酸序列）。小的 *ktup* 值搜索敏感性高，但是需要更多时间。

> University of Virginia 的 William Pearson 在线提供 FASTA 程序。可以访问 http://alpha10.bioch.virginia.edu/fasta/。另一个试用 FASTA 的地址是欧洲生物信息研究所的网址（http://www.ebi.ac.uk/fasta33/）。
> 从 NCBI 的主网页的顶部栏上点击 BLAST（或者直接访问 http://www.ncbi.nlm.nih.gov/BLAST/）。选择“BLAST 2 Sequences”。

3.3.4 局部比对搜索工具（BLAST）

BLAST 是一个局部比对搜索工具，用来确定一条查询序列和一个数据库的比对，但不引入间隙（Altschul 等，1990）。现在可用的 BLAST 版本已允许比对中引入间隙。我们将在第 4 章详细介绍 BLAST，那里我们将描述启发式算法。这里我们介绍“BLAST 2 Sequences”的实际使用，它是 NCBI 上 BLAST 网址的一部分，允许对两条序列执行局部双序列比对（Tatusova 和 Madden，1999）。使用这个程序执行蛋白质（或 DNA 序列）的双序列比对非常容易。执行以下步骤。

① 如果想比对两 DNA 序列，选择程序 blastn（指“BLAST nucleotides”）。如果是蛋白质，选择 blastp（指“BLAST proteins”）。我们的例子中为蛋白质。

② 输入序列或者它们的 accession numbers。例如，人 RBP4 的 accession number（NP_006735）和牛 β-乳球蛋白（P02754）（图 3-27）。

③ 选择可选参数：

a. 可以从 6 个打分矩阵中选择 BLOSUM62、BLOSUM50、BLOSUM90、PAM250、PAM70 和 PAM30。选择 PAM250。

b. 可以改变间隙产生和延长的罚分。

c. 对于 blastn 搜索，可以改变奖励和罚分值。

d. 还有其他参数可以改变，例如词的大小（word size）、期望值（expect value）、过滤（filtering）和截断值（dropoff value）。我们在第 4 章会有详细的讨论。

④ 点击“align”。一个典型的结果如图 3-28 所示。注意对于像例子中的远相关蛋白，如果选择了合适的打分矩阵（PAM250）和间隙罚分，只有一个比对返回。输出包括了一张点阵图（dot plot），图中两序列比对的部分沿对角线表示（框 3-7）。输出还包括如在图 3-5 中所见到的氨基酸单字母编码的双序列比对。

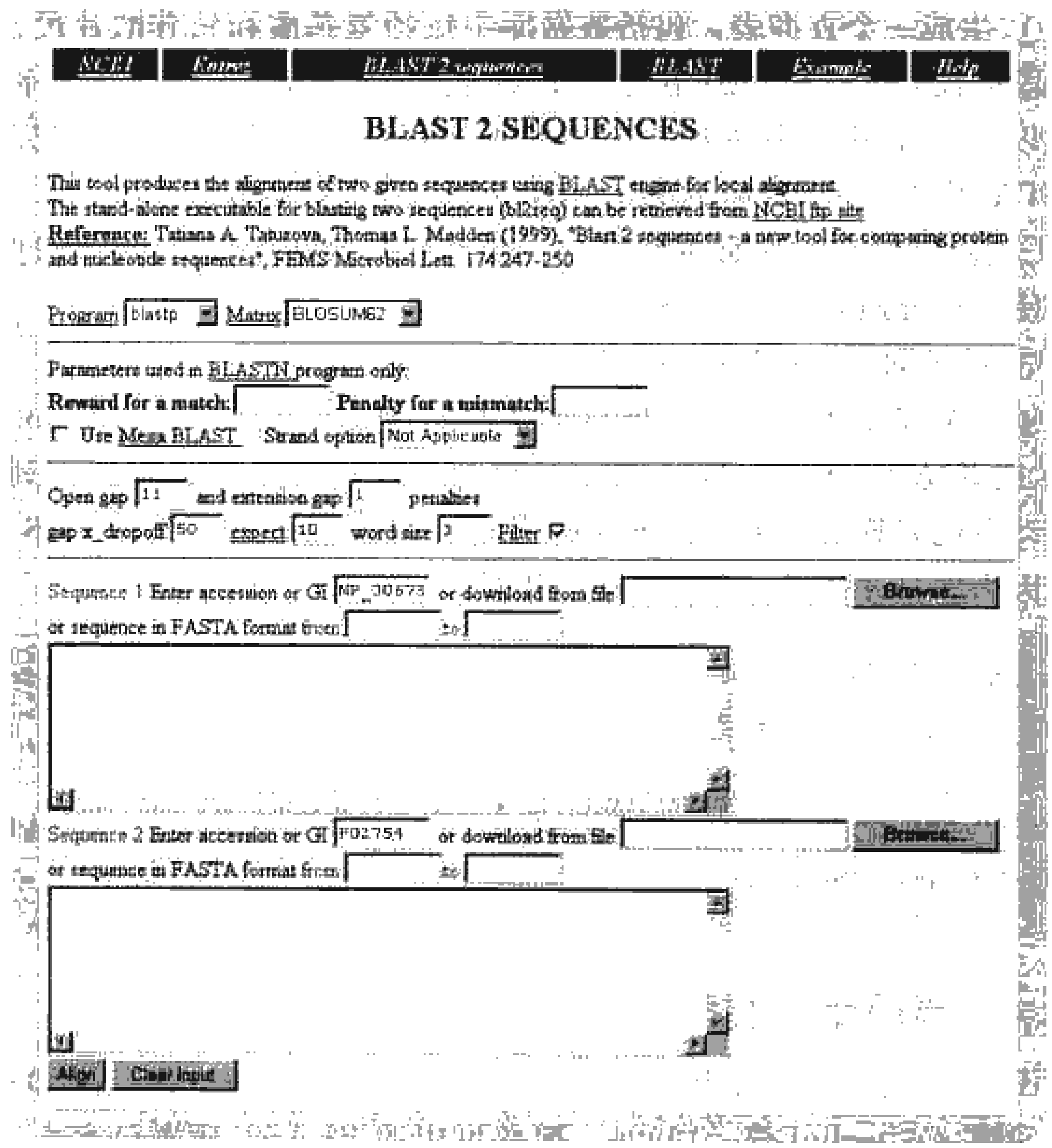


图 3-27 使用 NCBI BLAST 2 Sequences 算法进行序列比对的一个例子。输入人 RBP4 的 accession number（NP _ 006735）和牛 β-乳球蛋白（P02754）。①选择合适的程序（蛋白质搜索使用 blastp，DNA 搜索使用 blastn），②输入 FASTA 格式的序列或则 accession number（如本例），③选择可选参数，例如间隙罚分，④点击“Align”

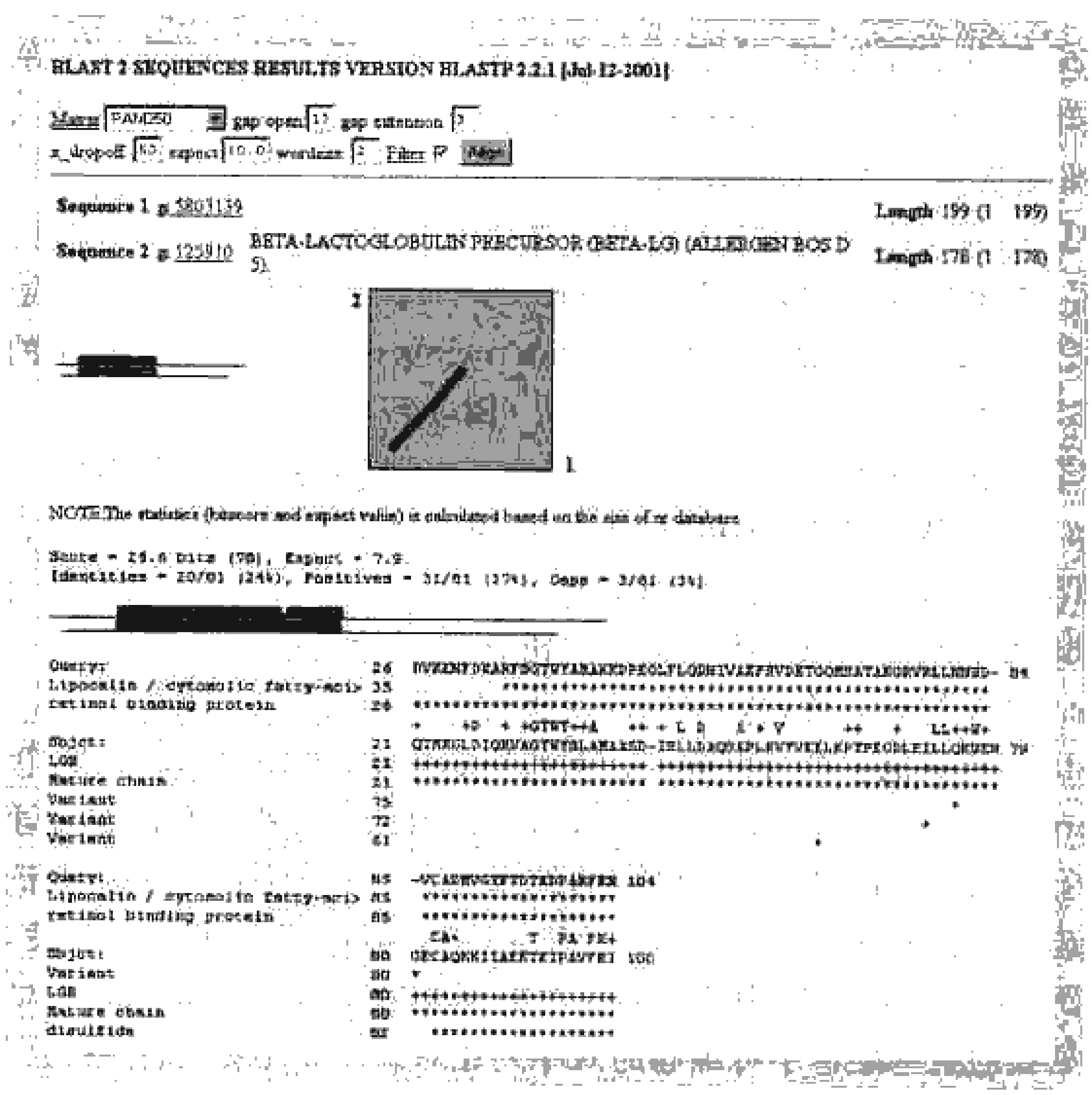


图 3-28 BLAST 2 Sequences 程序输出的一个例子。注意对于这些远相关蛋白，PAM 矩阵设为 PAM250，间隙引入和延长的罚分调整为 13 和 3。使用 PAM30、PAM70 和 BLOSUM 矩阵检测不到显著的比对。输出包括一概要显示双序列比对的点阵图和比对的氨基酸序列。注意输出还包括对位分值（bit score）、期望值和一致性、正确性（positives）以及间隙的描述，我们在第 4 章会讨论这些输出。

框 3-7　点阵图（dot plot）

点阵图或点阵是比较两条序列的图形化方法。一条蛋白质或核苷酸序列沿 x 轴放置，另一条沿 y 轴放置。氨基酸相同的位置放一个点。两条序列中相同的区域会形成一条对角线。可能会有多条对角线。或者也可以调整参数，让两条序列一组残基（如一行的 5 个核苷酸）匹配才放置一个点。如果蛋白质（或核苷酸）序列用点阵图进行自我比较自身，序列内的重复区域可以容易地显示出来。

我们在第 13 章比较细菌基因组序列时也会碰到点阵图，在第 15 章（真菌）中也能见到。两酿酒酵母染色体序列使用 BLAST 相互比对。结果的点阵图出现许多对角线，表示同源区域。这个点阵图使研究者提出整个酵母基因组大概在 1 亿年前发生复制。

Dotler 是一个基于 web 的对角线画图工具，可以从瑞士生物信息研究所的网站得到（http://www.isrec.isb-sib.ch/java/dotler/Dotler.html）。

3.4　双序列比对的显著性：一致性百分比

我们怎样才能判断两条序列从进化观点上来说是显著相关的？RBP 和 β-乳球蛋白的比对的氨基酸一致性为 26%，这个值随机情况下也能发生吗？

一个经验规则就是如果两条序列在 150 个氨基酸或者更长的范围内有 25%或以上的氨基酸一致性，那么它们可能是显著相关的（Brenner 等，1998）。考虑有 70 个氨基酸长度的比对，如果它们共享 25%的氨基酸一致性，那么认为这两条序列“显著相关”，这种做法比较普遍。但是，Brenner 等（1998）认为这可能会产生错误，部分是因为当今的分子序列数据库变得越来越大，使得随机找到这样比对的可能性增大。对于有 70 个氨基酸残基的比对，40%氨基酸一致性是一个认为两蛋白同源的合理阈值（Brenner 等，1998）。如果两蛋白在一个适当长度（如 70～100 个氨基酸残基）的片段上有 20%～25%的一致性，并且它们刚好在“模糊区域”（图 3-19），则确认它们的同源性就比较困难。如两蛋白比对时只有 10%～20%的一致性，那么可以认为它们是完全不相关的。这种情况特别发生在插入间隙以后可以提高任意两个序列比对的分值时。

在第 10 章我们将发现多序列比对的敏感性比双序列比对高 3 倍。我们在第 5 章也会看到序列比对的打分矩阵（“profiles”）可以被定制，能大幅度提高搜索的敏感性。

3.4.1　双序列比对统计显著性检验

如果两蛋白质只有有限的氨基酸一致性（例如 20%～25%），我们怎么才能确定它们是否显著相关？比对算法提供了双序列比对的分值，或者是查询序列和整个数据库最佳比对的分值（第 4 章）。我们需要进行统计检验来决定匹配是真阳性（即两比对的蛋白是真正同源的）还是假阳性（即它们是由于算法原因而随机比对在一起的）（图 3-29）。对于算法没有提供的比对，例如因为它们的分值在某个阈值以下，我们需要估计这些序列是否是真阴性（即真的不相关）或者它们是假阴性（即它们是同源序列，但得到的分值显示它们不是同源序列）。

这样一个比对算法的主要目的就是最大化序列比对的敏感性和特异性（图 3-29）。敏感性的计算是以真阳性的数目除以真阳性和假阴性数目之和。这是衡量一个算法能正确确定真实相关序列的一个尺度。特异性的计算是以真阴性的数目除以真阴性和假阳性数目之和。这个值描述了非同源序列的比对。

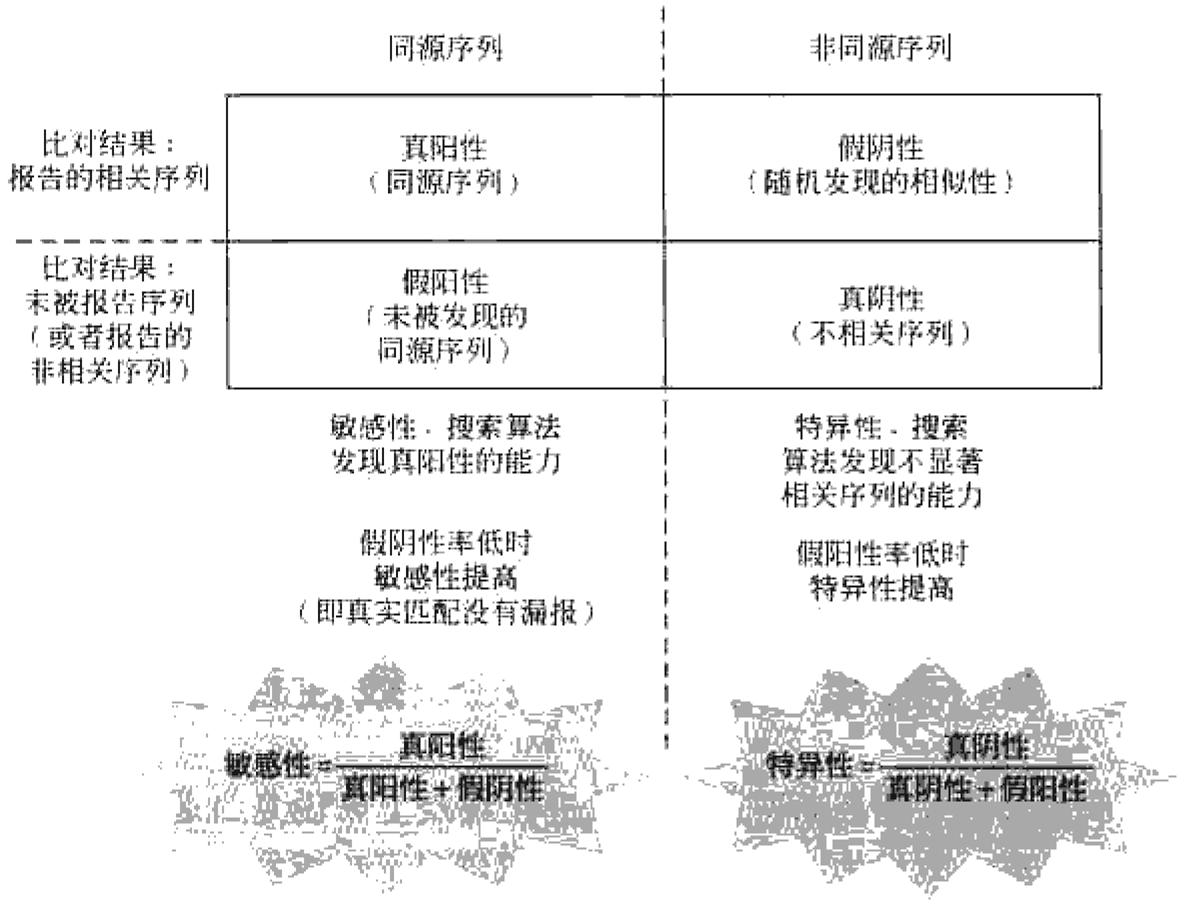


图 3-29 序列比对、双序列比对（本章）或数据库搜索（第4章）的结果是真阴性还是假阴性和阳性？比对的统计分析可以允许估计比对是否是真阳性（即是同源序列的比对）。理想的比对算法能最大化敏感性和特异性

决定比对是否具有统计显著性有几个主要的方法。如一个蛋白（如RBP4）能和其他蛋白比较（如β-乳球蛋白），我们可以使用打分系统产生一个原始分值。为了评估比对的显著性，我们可以把这个结果和RBP与其他序列比对的结果比较，例如：

① 许多其他已知的不和RBP4同源的蛋白质；

② 其他序列（例如β-乳球蛋白自身）可以被打乱，同时保持一些属性不变（如长度、氨基酸组成）；

③ 随机产生的序列。

我们将考虑先全局比对然后局部比对来研究这个问题。在第4章，我们将把统计检验应用到使用工具（如BLAST搜索数据库）上。

3.4.2 全局比对的统计显著性

我们对全局比对的统计特性所知甚少（相对于局部比对而言），但是我们可以考虑如何去评估它们的显著性。让我们以RBP4和β-乳球蛋白为例。我们首先对它们比对，得到分值，然后打乱β-乳球蛋白序列100次，执行100次比对，记录下这些分值，然后以真实分值去比较这些随机分值的均值。我们用随机分值的标准偏差来表示真实分值和随机分值平均值的差：

$$Z=\frac{x-\mu}{\sigma} \tag{3-3}$$

式中，x 指两序列比对的真实分值；μ 为随机打乱序列比对的平均分值；σ 为这些值的标准偏差。

考虑RBP和数据库中100万个蛋白序列比较的情况，而不是仅和β-乳球蛋白比较。这种情况下，设置 α 水平（拒绝无效假设的概率水平）在传统的 $p<0.05$ 上已不合适。相反地，α 必须除以实验的数目，这个例子中，你定义的概率水平必须 $<0.05/10^6$（或 5×10^{-8}）才能具有显著性。这个称作Bonferroni修正（见第7章）。即使比对没有统计显著性，从双序列比对得到一个大约Z值也是有可能的（例如10，表示这个分值在均值之上的10倍标准偏差处）。

我们可以从 Z 值得到什么结论？如果随机打乱的 100 个蛋白质序列的比对分值都小于 RBP4 和 β-乳球蛋白比对的真实分值，这表示随机发生的概率值（p）<0.01（这样我们可以拒绝 RBP4 和 β-乳球蛋白序列无显著相关的无效假设）。

我们可以使用计算机算法（如 GAP，来自 GCG 软件包）来进行这种随机打乱测试。GAP 计算全局双序列比对的分值，这个例子中为 37（图 3-24）。这个程序也执行其他许多次比对，这个例子中为 RBP 和 100 个打乱的 β-乳球蛋白序列的比对。例子中，100 个随机 β-乳球蛋白序列得到的平均分值为 8.7±4.2（图 3-30）。这样，计算 Z 值为 $(37-8.7)/4.2=6.7$。

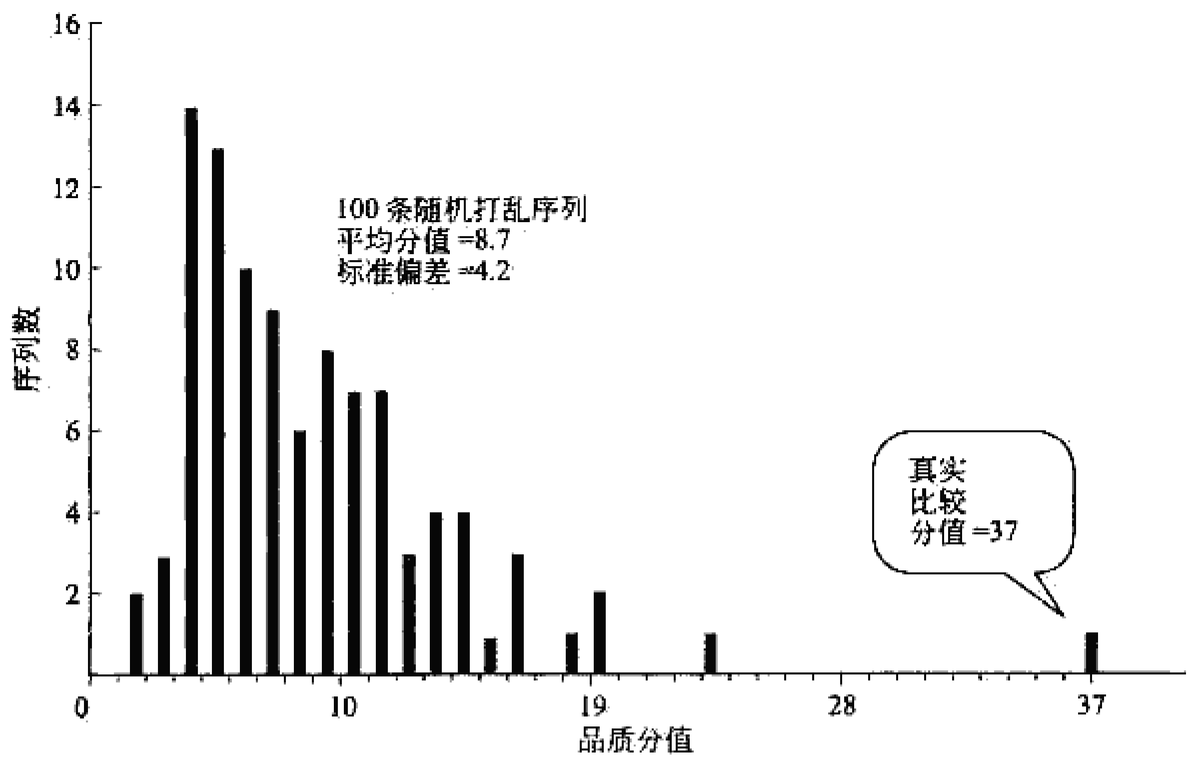


图 3-30 使用随机测试评估双序列比对的统计显著性。使用 GCG 的 GAP 程序对人 RBP4 和 β-乳球蛋白比对。比对产生的分值为 37。然后 β-乳球蛋白序列被随机打乱 100 次，保持氨基酸长度和组成不变。每个分值都表示在图上。打乱序列的均值和标准偏差为 8.7±4.2。假设随机值服从正态分布，则可以把 Z 值转换成随机比较的可能值

我们希望能使 Z 分值和概率值相关，从而来衡量统计显著性。但是，要正确地做这件事不可能，除非我们知道这些随机值的整体分布。对于全局比对，这个分布还没有被很好地研究，这样，Z 分值不能转换成概率值。对于局部序列比对，统计理论的发展要远好于此。

3.4.3 局部比对的统计显著性

许多数据库搜索程序（如 BLAST，第 4 章）依赖于局部比对。而且，许多双序列比对程序用局部比对比较两条序列。局部比对在一起的区域称作高分片段对（high-scoring segment pairs）或者 HSPs。与 BLAST 相关的最重要的统计量为期望值（E 值），它描述了搜索给定大小的数据库随机情况下期望得到的匹配数目。E 值是衡量比对结果统计显著性的一个尺度，我们将在第 4 章中详细讨论。

3.5 展望

DNA 和蛋白质序列的双序列比对是生物信息学的最基本操作之一。双序列比对允许人们确定任意两序列之间的相关性，由其确定的相关程度可以帮助人们考虑它们是否有同源性（来自于一共同进化祖先）。本书的余下部分几个所有主题都依赖于序列比对。在第 4 章，我们将介绍用查询序列去搜索巨大的 DNA 和/或蛋白质数据库。数据库搜索一般包括大量的局部序列比对，返回的结果序列按相关性程度从高至低排列。

用于双序列比对的算法开发于 19 世纪 70 年代，开始于由 Needleman 和 Wunsch（1970）提出的全局比对算法。Dayhoff（1978）提出了 PAM 打分矩阵，允许比较和评估两

条远相关的分子序列。至19世纪80年代，提出了局部比对算法［特别参见Seller（1974）、Smith和Waterman（1981）以及Smith等（1981）的工作］。实际上，当今使用的双序列比对程序只是有限的几组软件，大多数是可以通过因特网免费获得的。

现有的双序列比对算法的敏感性和特异性还将被继续评估。近来，双序列比对领域继续发展，包括屏蔽低复杂度序列的方法（第4章会有讨论）和比对中间隙罚分的理论模型。

3.6 常见问题

执行双序列比对通常发生的错误是没有得到合适的输入序列：特别是比较DNA和蛋白质序列（或者相反）是不可能的。

双序列比对算法的可选参数的选择可以很大程度影响比对的结果。如果使用BLAST 2 Seuqnces并以缺省参数比较人RBP4和牛β-乳球蛋白，那么检测不到匹配。

任意两条序列可以比对在一起，即使它们是不相关的。在一些情况下，即使两条蛋白质在100个氨基酸长度的片段上有>30%的氨基酸一致性，它们还是不同源（进化相关）的。评价序列比对的生物学显著性总是重要的。这可以用于寻找细胞功能、基本结构是否相同，或三维结构是否相似的证据。

3.7 网络资源

双序列比对可以使用实现了全局或局部比对算法的软件包来执行。所有情况下，两条蛋白质或者核苷酸序列直接比较。

许多网址提供基于全局比对（表3-5）或局部（表3-6）比对的免费双序列比对算法。这些网站包括NCBI的BLAST 2 sequences、Baylor College of Medicine（BCM）launcher、ExPASy上的SIM程序，以及位于Georgetown University的Protein Information Resource（PIR）的SSEARCH。

表3-5 全局双序列比对算法

资源	描述	URL
ALIGN	位于法国GENESTREAM服务器	http://www2.igh.cnrs.fr/bin/align-guess.cgi
GAP	来自Genetics Computer Group (GCG)	http://www.gcg.com
Needle	来自Institut Pasteur，实现了Needleman-Wunsch全局比对	http://bioweb.pasteur.fr/docs/EMBOSS/needle.html
Pairwise alignment (various)	来自University of Virginia(Bill Pearson)	http://alpha10.bioch.virginia.edu/fasta/
Pairwise	双序列比对工具(全局或局部)	http://informagen.com/Applets/Pairwise/
Pairwise Sequence alignment	来自Baylor College of Medicine	http://searchlauncher.bcm.tmc.edu/
Stretcher	来自Institut Pasteur，全局比对	http://bioweb.pasteur.fr/docs/EMBOSS/stretcher.html
Vector NTI Suite 7	来自Informax	http://www.informaxinc.com/

注：EMBOSS为The European Molecular Biology Open Software Suite（http://www.uk.embnet.org/Software/EMBOSS）的缩写。

表 3-6 局部比对算法

资 源	描 述	URL
BestFit	来自 Genetics Computer Group (GCG)	http://www.gcg.com
BLAST 2 Sequences	位于 NCBI	http://www.ncbi.nlm.nih.gov/BLAST/
Est2genome	来自 Institut Pasteur,比对表达序列标签(EST)和基因组 DNA	http://bioweb.pasteur.fr/docs/EMBOSS/est 2 genome.thml
LALIGN	在两序列中寻找多个匹配子段	http://www.ch.embnet.org/software/LALIGN_form.html
Pairwise alignment (various)	来自 University of Virginia(Bill Pearson)	http://alpha10.bioch.virginia.edu/fasta/
Pairwise	双序列比对工具(全局和局部)	http://informagen.com/Applets/Pairwise/
Pairwise Sequence Alignment	来自 Baylor College of Medicine	http://searchlauncher.bcm.tmc.ecu/
PRSS	来自 University of Virginia(Bill Pearson)	http://fasta.bioch.virginia.edu/fasta/prss.htm
SIM	来自 ExPASy 的蛋白质序列比对工具	http://www.expasy.ch/tools/sim prot.html
SIM	SIM,gap,位于 Department of Computer Science,Michigan Tech	http://genomes.cs.mtu.edu/align.html
SSEARCH	来自 Protein Information Resource	http://pir.georgetown.edu/pirwww/
Vector NTI Suite 7	来自 Informax	http://www.informaxinc.com/

注：EMBOSS 为 The European Molecular Biology Open Software Suite (http://www.uk.embnet.org/Software/EMBOSS) 的缩写。

现有许多商业软件包，包括 Genetics Computer Group (GCG) 软件包和 Vector NTI (Informax)。GCG 软件包包括 GAP 程序（用于全局双序列比对）和 BESTFIT（用于局部比对）。在基于 Unix 的版本中，命令"gap rbp.pep lactoglobulin.pep-ran=100"将运行 gap 全局比对算法，并估计比对的统计显著性（使用 100 条打乱的 β-乳球蛋白序列）。也有基于 web 的界面 SeqWeb。

3.8 问题讨论

1. 如果你希望比较任意两条序列，是否存在一个"正确"的打分矩阵供你选择？如何知道哪个打分矩阵是最合适的？

2. 许多蛋白质（或 DNA）序列有多个分开的结构域（我们在第 8 章讨论结构域）。考虑一个有两个结构域的蛋白，其中一个进化较快，另一个进化较慢。它和另一条蛋白（或 DNA）双序列比对时，你打算使用打分矩阵 PAM40 和 PAM250 来执行两个分离的比对还是选择一个"中间的"矩阵？为什么？

3. 在 Margaret Dayhoff 和其同事发表蛋白质数据集（protein atlas）和打分矩阵的 13 年前，Emile Zuckerkandl 和 Linus Pauling (1965) 为当时所有的数十个球蛋白序列提供了一打分矩阵（图 3-31）。图中的行（y 轴）显示了原始球蛋白氨基酸，列显示了观察到的替换。如果至少 20%的位置发生了替换，那么该单元格就被填上一个数值。注意深灰色单元格，这些位置的氨基酸替代没有被观察到。灰色阴影单元格的氨基酸替代是非常保守的。

该矩阵的数据和 Dayhoff 及其同事描述的矩阵比较有什么异同？哪些替换发生最少，而哪些发生最频繁？如果现在让你填这张表，你会如何？

替代残基（替代发生的残基位点百分比）

序列（原氨基酸）	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A				28			31	33								31				
R									50			58				25				
N	33			47					33			33				33	33			
D	44		22				47	34	22			28				25				
C	(66)																			
Q				56			30		40			70								
E	50			44				38				41			24					
G	51			33			30					27				36				
H				26							26	30				22	22			
I	39										58									46
L	21									23		23		28						30
K	23	21		28			31	23			21					21				
M	22									22	89			22						45
F									22		61									
P	50			43			57	43				21								
S	49			24			24	36				24					40			
T	32						28	24				24				52				
W	(40)										(40)			(60)						
Y									(33)					(50)						
V	36									21	43	21								

图 3-31 球蛋白的替换频率（来自 Zuckerkandl 和 Pauling，1965）。氨基酸以它们对应的三字码的字母顺序排列，行对应于来自人、其他灵长类、马、牛、猪、七鳃鳗和鲤鱼的数十条血红蛋白和肌球蛋白序列的比对中最初的氨基酸。数字代表在残基位置给定替换发生的百分比。例如，观察到 33% 的位置的丙氨酸被甘氨酸替换。深灰色方框内的替换没有被观察到。较小频度的替换（百分比<20%）以白色格子表示（数值没有给出）。"非常保守"的替换（百分比≥40%）表示在灰色方框中。例如，89% 的甲硫氨酸的位置被替换成亮氨酸。相同的替代以黑色方框表示。括号内值表示样本量很小，意味着由这些数据得出结论必须小心。使用经授权

Joshua Lederberg 帮助 Zuckerkandl 和 Pauling（1965）制作了这个矩阵。他们使用 IBM7090 计算机，这是基于晶体管技术的第一台商业计算机，价值 300 万美元左右。它的内存为 32768 个双字（binary word）或者大约 131000 字节（阅读关于 Lederberg 获得 1958 年诺贝尔奖的资料，参考 http://www.nobel.se/medicine/laureates/1958/index.html）。

3.9 问题

1. 病毒逆转录酶（例如 *pol* 基因产物由 HIV-1 编码）具有人类同源物。HIV-1 逆转录酶的 GenBank accession number 为 NP_057849（使用 Entrez 确认这个 accession number 是否正确）。Entrez 搜索会列出许多人类病毒相关基因产物，包括 874 个氨基酸残基长度的逆转录病毒相关 pol 多聚蛋白（P10266）。使用 BLAST 2 Sequences 中的 blastp 对这两个蛋白进行双序列比对。

这个搜索的缺省设置包括使用 BLOSUM62 打分矩阵。结果显示于图 3-32 和表 3-7。期望值大约为 1×10^{-61}，表示两蛋白紧密相关，即使它们在 919 个氨基酸长度上仅有 25%的氨基酸一致性。在表中填上使用 BLOSUM62、BLOSOM50 和 BLOSUM90 打分矩阵的搜索结果。改变搜索参数的效果是什么？

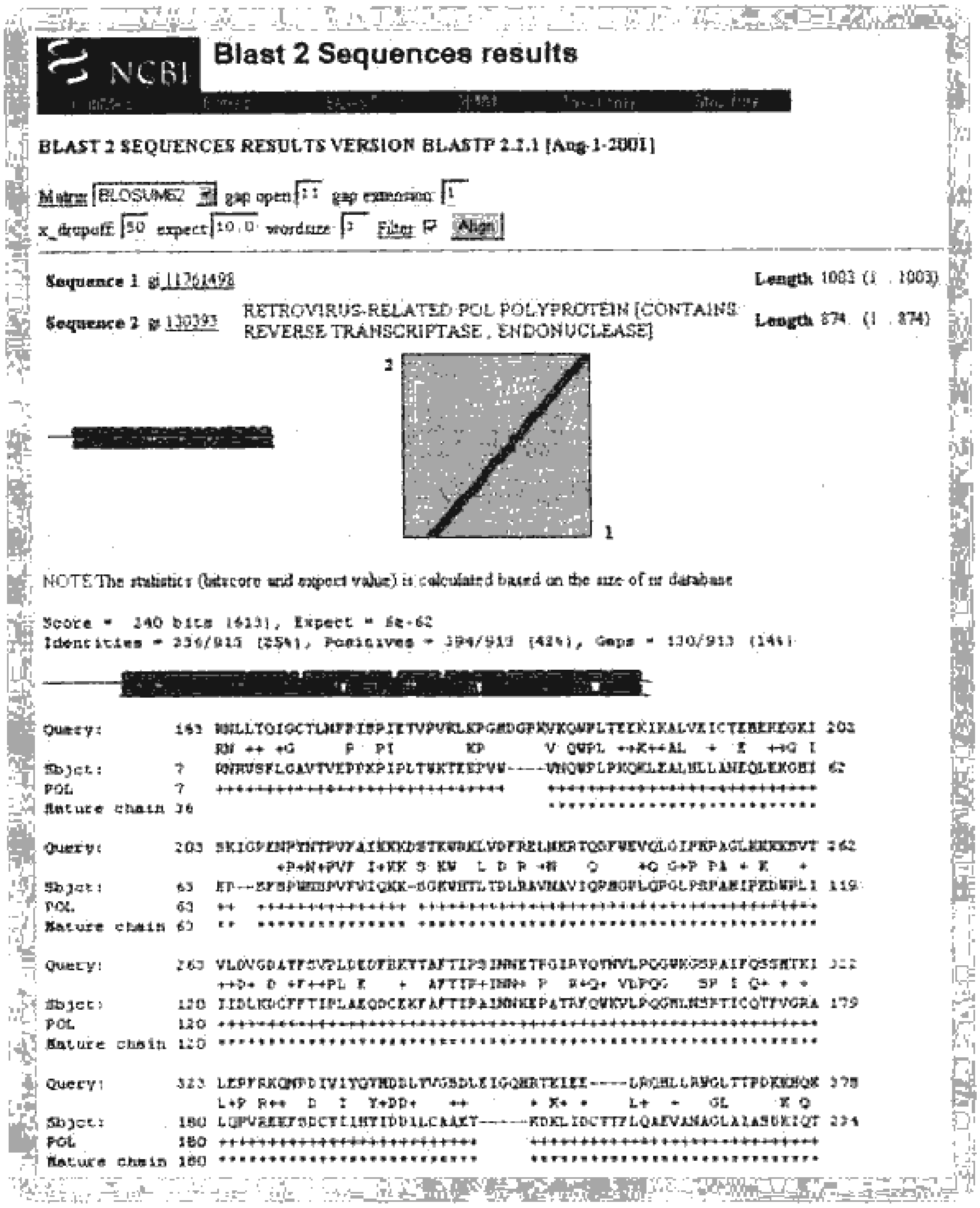


图 3-32　HIV-1 pol(NP_057849) 和人逆转录酶 (P10266) 双序列比对的部分输出。两个蛋白总体上有 25%的氨基酸一致性

表 3-7　改变打分矩阵对 HIV-1 pol 和相关人类蛋白双序列比对的效应

打分矩阵	分　值	期望值	相同残基	相似残基	间隙罚分	间隙个数
BLOSOM50						
BLOSOM62	240bits (613)	1×10^{-61}	236/913 (25%)	394/913 (43%)	GO=11, GE=1	130/913 (14%)
BLOSOM90						
PAM30						
PAM70						
PAM250						

注：GO 代表 gap opening，GE 代表 gap extension。

2. 下一步使用 PAM30、PAM70 和 PAM250 矩阵对问题 1 的两蛋白进行双序列比对。它们的期望值多大？比对在一起的氨基酸有多少？搜索结果和使用 BLOSUM 矩阵的结果有何不同？

3. 使用 BLAST 2 Sequences（或者其他列在表 3-5 中的程序）对 RBP 和 β-乳球蛋白进

行局部双序列比对。使用低的间隙罚分，然后再重复一次，结果是什么？

4. 比较现代人线粒体 DNA 和已绝种的尼安德特人 DNA（Neanderthal DNA）。首先获取线粒体编码细胞色素氧化酶（cytochrome oxidase）基因的核苷酸序列（搜索 NCBI 的 taxonomy 部分，网址 http://www.ncbi.nih.gov/taxonomy，然后选择"extinct organisms"寻找尼安德特人 DNA），然后执行双序列比对，记录核苷酸一致性的百分比。

5. 我们已经看到一些基因家族变化较慢（如图 3-8 的 GAPDH），而一些变化较快（见框 3-3）。对于问题 3 中的细胞色素氧化酶基因，你如何确定它的相对变化快慢？尝试使用 BLAST 2 Sequences。

6. Aphrodisin 和 odorant-binding protein（气味结合蛋白）都是 lipocalin。首先得到啮齿类中这两个蛋白的 accession number，然后执行双序列比对（使用 BLAST 2 Sequences）。记录氨基酸一致性百分比、相似百分比、期望值和位分值。衡量它们的相关性，哪个矩阵最有帮助？

7. 海龟和鸟类中哪个与鳄鱼的相关性比较近？首先，尝试用 PubMed 寻找答案；然后选择一个基因或蛋白，使用 BLAST 2 Sequnces 比对。

8. 腔棘鱼类 *Latimeria chalumnae* 是圆鳍鱼类，它们被称为"活化石"。人们很早就认为它们灭绝于至少 9000 万年以前，现在在海洋中发现几例标本。令人惊奇的是，对线粒体序列的系统进化分析表明腔棘鱼类比鲱鱼类更接近于人类（Lewin，2001）。在人类、腔棘鱼类和鲱鱼类的线粒体中找一些 DNA 序列，并得到它们的 accession number，然后执行双序列比对，看看是否真是如此。提示：使用 PubMed 去寻找生物的属名和种名，鲱鱼是 *Clupea harengus*。下一步使用种名在 Entrez 中搜索，例如"*Clupea harengus* Mitochondrion"。

9. PAM1 矩阵（图 3-11）是非互易的：例如丙氨酸到精氨酸的变化概率不等于从精氨酸到丙氨酸的变化概率。为什么？

10. 河马与猪、鲸中谁的相关性更近？为回答这个问题，首先从每种生物中找到血红蛋白序列，然后进行双序列比对，记录氨基酸一致性百分比。

3.10 自测题

1. 匹配下列氨基酸和它们的单字母代码

天冬酰胺	Q
谷氨酰胺	W
色氨酸	Y
酪氨酸	N
苯丙氨酸	F

2. 直系同源物定义为

(a) 不同物种中具有共同祖先的同源序列；

(b) 具有较小的氨基酸一致性但是有较大的结构相似性的同源序列；

(c) 同一物种中由基因复制产生的同源序列；

(d) 同一物种中具有相似的并且通常是冗余的功能的同源序列

3. 根据 PAM 打分矩阵，下列哪个氨基酸最不容易突变？

(a) 丙氨酸；

(b) 谷氨酰胺；

(c) 甲硫氨酸；

(d) 半胱氨酸

4. PAM250 矩阵定义的进化距离为两同源序列在给定的时间有多少百分比的氨基酸发生改变？

(a) 1%；

(b) 20%；

(c) 80%；

(d) 250%

5. 下列哪个句子最好地描述了两序列全局比对和局部比对的不同?

(a) 全局比对通常用于比对 DNA 序列，而局部比对通常用于比对蛋白质序列；

(b) 全局比对允许间隙，而局部比对不允许间隙；

(c) 全局比对寻找全局最大化，而局部比对寻找局部最大化；

(d) 全局比对比对整条序列，而局部比对寻找最佳匹配子序列

6. 假如你有两条远相关的蛋白。为了比较它们，最好使用下列哪个 BLOSUM 和 PAM 矩阵?

(a) BLOSUM45 或 PAM250；

(b) BLOSUM45 或 PAM1；

(c) BLOSUM80 或 PAM250；

(d) BLOSUM10 或 PAM1

7. 与 PAM 打分矩阵比较，BLOSUM 打分矩阵的最大区别在哪里?

(a) 它最好用于比对相关性很近的蛋白；

(b) 它是基于近相关蛋白的全局多序列比对；

(c) 它基于远相关蛋白的局部多序列比对；

(d) 它结合了局部和全局比对的信息

8. 正确或错误：具有 30%氨基酸一致性的两蛋白质有 30%的同源性。

9. 全局比对算法（如 Needleman-Wunsch 算法）保证能找到最佳比对。这样的一种算法

(a) 把两条比较的蛋白质放到一个矩阵中，然后通过穷尽搜索每一个可能的比对组合来找到最佳的分值；

(b) 把两条比较的蛋白质放到一个矩阵中，然后通过迭代递归的方法来找到最佳的分值；

(c) 把两条比较的蛋白质放到一个矩阵中，然后通过寻找最佳子序列的方法来找到最佳的比对；

(d) 能用于蛋白质，但不能用于 DNA 序列

10. 数据库搜索中或双序列比对中，敏感性定义为

(a) 搜索算法寻找真阳性（即同源序列）和避免假阳性（即不相干序列，但具有高的相似分值）的能力；

(b) 搜索算法寻找真阳性（即同源序列）和避免假阳性（即没有被搜索算法报告的同源序列）的能力；

(c) 搜索算法寻找真阳性（即同源序列）和避免假阴性（即不相干序列，但具有高的相似分值）的能力；

(d) 搜索算法寻找真阳性（即同源序列）和避免假阴性（即没有被搜索算法报告的同源序列）的能力

3.11 推荐读物

我们以介绍同源性的概念为开始介绍本章，同源性这个词经常被误用。一篇 Reeck 等（1987）的一页长的文章提供了同源性和相似性的权威的和标准的定义。最近的关于同源性和系统发生的讨论可参见 Tautz（1998）。

所有描述序列比对的文章都在一个分子进化的模型中考虑两同源序列的趋异程度。Russell F. Doolittle（1981）提供了清晰的、深入的关于序列比对的综述。William Pearson（1996）概述了序列比对。他描述了相似分值、敏感性、选择性和搜索程序（如 Smith-Waterman 和 FASTA）的统计学。其他的关于双序列比对的综述包含在“Trends Guide to Bioinformatics”中由 Altschul（1998）和 Brenner（1998）写的短文章。

关于双序列比对算法的研究，一个重要的文献是 Margaret O. Dayhoff 和其同事 1978 年出版的一部著作（Dayhoff，1978）。这部著作的一大部分是蛋白质序列数据集和相伴的进化系统的构建。该书第 22 章介绍了可接受点突变（accepted point mutation）的概念，第 23

章描述了各种 PAM 矩阵。至 20 世纪 90 年代早期，更多的蛋白质序列出现，Steven 和 Jorja Henikoff（1992）提出了 BLOSUM 矩阵。这篇文章对使用这些打分矩阵做了非常好的介绍，特别是比较了 PAM 矩阵和 BLOSUM 矩阵的性能。在本书的以后部分（第 4 章和第 5 章）中，我们会在数据库搜索中广泛使用这些矩阵。

全局比对的算法最初由 Needleman 和 Wunsch（1970）提出，后来局部比对的算法由 Smith 和 Waterman（1981）、Waterman 和 Fitch（1981）引入。双序列比对的敏感性（检测远相关序列的能力）和选择性（避免不相关序列）的问题由 Pearson 和 Lipman 在 1981 年的一篇介绍 FASTA 程序的文章中提出。

最近多篇文章讨论了序列打分统计学的问题。瑞士生物信息学院的 Marco Pagni 和 C. Victor Jongeneel（2001）提供了一个很好的综述，包括对 BLAST 打分的统计学的讨论，这和第 4 章和第 5 章中的内容相关。

最后，Steven Brenner、Cyrus Chothia 和 Tim Hubbard（1998）比较了数种双序列比对方法。这篇文章很值得推荐用于学习如何评估各种不同的算法（例如，我们在第 10 章的多序列比对中会见到类似的方法）。阅读这篇文章能帮助你理解为什么统计分值比其他的搜索参数（如原始分值或者一致性百分比）在解释双序列比对的结果上更有效。

参 考 文 献

Altschul, S. F. Fundamentals of database searching. *Bioinformatics: A Trends Guide* **1998**, 7–9 (1998).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

Anfinsen, C. *The Molecular Basis of Evolution*. John Wiley & Sons, Inc., New York, 1959.

Brenner, S. E. Practical database searching. *Bioinformatics: A Trends Guide* **1998**, 9–12 (1998).

Brenner, S. E., Chothia, C., and Hubbard, T. J. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA* **95**, 6073–6078 (1998).

Chothia, C., and Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *Embo J.* **5**, 823–826 (1986).

Dayhoff, M. O. *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, MD, 1978.

Doolittle, R. F. Similar amino acid sequences: Chance or common ancestry? *Science* **214**, 149–159 (1981).

Doolittle, R. F. *OF URFS AND ORFS: A Primer on How to Analyze Derived Amino Acid Sequences*. University of Science Books, Mill Valley, CA, 1987.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. *Biological Sequence Analysis*. Cambridge University Press, Cambridge, 2000.

Gonnet, G. H., Cohen, M. A., and Benner, S. A. Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443–1445 (1992).

Gotoh, O. An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705–708 (1982).

Henikoff, S., and Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919 (1992).

Henikoff, J. G., and Henikoff, S. Blocks database and its applications. *Methods Enzymol.* **266**, 88–105 (1996).

Lewin, R. A. Why rename things? *Nature* **410**, 637 (2001).

Myers, E. W., and Miller, W. Optimal alignments in linear space. *Comput. Appl. Biosci.* **4**, 11–17 (1988).

Needleman, S. B., and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).

Newcomer, M. E., et al. The three-dimensional structure of retinol-binding protein. *EMBO J.* **3**, 1451–1454 (1984).

Pagni, M., and Jongeneel, C. V. Making sense of score statistics for sequence alignments. *Brief Bioinform.* **2**, 51–67 (2001).

Pearson, W. R. Effective protein sequence comparison. *Methods Enzymol.* **266**, 227–258 (1996).

Pearson, W. R., and Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448 (1988).

Reeck, G. R., et al. "Homology" in proteins and nucleic acids: A terminology muddle and a way out of it. *Cell* **50**, 667 (1987).

Sellers, P. H. On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.* **26**, 787–793 (1974).

Smith, T. F., and Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).

Smith, T. F., Waterman, M. S., and Fitch, W. M. Comparative biosequence metrics. *J. Mol. Evol.* **18**, 38–46 (1981).

Tatusova, T. A., and Madden, T. L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174**, 247–250 (1999).

Tautz, D. Evolutionary biology. Debatable homologies. *Nature* **395**, 17, 19 (1998).

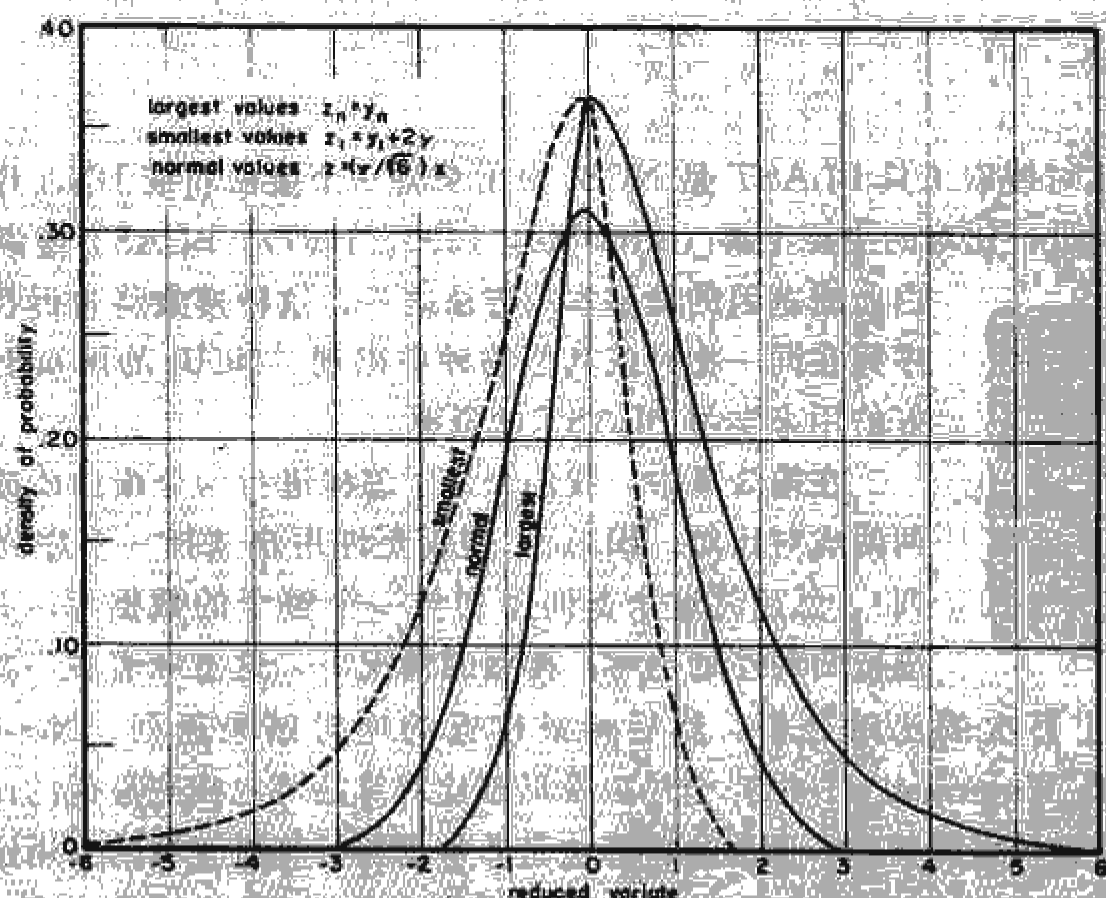
Zuckerkandl, E., and Pauling, L. Evolutionary divergence and convergence in proteins. In: V. Bryson and H. J. Vogel (eds.), *Evolving Genes and Proteins*, Academic Press, New York, 1965, pp. 97–166.

180 FIRST ASYMPTOTIC DISTRIBUTION 5.2.7

of the variate, the first double exponential distribution has larger (smaller) densities than the normal one. The opposite is true for the second double exponential distribution.

Table 5.2.7. Selected Probabilities for Normal and Largest Values

Value	*Reduced Variate*		*Probabilities*		*Return Periods*	
	Largest	*Normal*	*Largest*	*Normal*	*Largest*	*Normal*
$\bar{x}-\sigma$	−.70533	−1	.13206	.15866	7.57	6.30
$\bar{x}+\sigma$	1.85977	1	.85581	.84134	6.93	6.30
$\bar{x}\pm\sigma$	—	—	.72375	.68268	—	—
$\bar{x}-2\sigma$	−1.98788	−2	.00068	.02275	1480.	43.96
$\bar{x}+2\sigma$	3.14232	2	.95773	.97725	23.7	43.96
$\bar{x}\pm2\sigma$	—	—	.95705	.95450	—	—
$\bar{x}-3\sigma$	−3.27043	−3	$3.7\cdot10^{-12}$	.00135	$.27\cdot10^{11}$	741
$\bar{x}+3\sigma$	4.42486	3	.98810	.99865	84.01	741
$\bar{x}\pm3\sigma$	—	—	.98810	.99730	—	—



Graph 5.2.7(1). Extreme and Normal Distributions

In Graph 5.2.7(2) the probabilities of the largest and the smallest values and the normal probabilities for the same mean and standard deviation

第 4 章描述了主要的数据库搜索工具 BLAST。BLAST 最初由 Altschul 等于 1990 年提出，对 BLAST 搜索得到的分值的统计解释是基于 20 世纪 50 年代发展的数学模型。在许多情况下，数值在群体中的分布假定服从正态（高斯）分布，如该图中所示（标记有“normal”的曲线）。但是，对一系列自然现象，数值的分布不是正态的。数据库搜索就是如此，即使用一条感兴趣的蛋白质或 DNA 序列在一个大的数据库中进行搜索，这也是我们在本章将描述的内容。最大分值的分布适合于极值分布（extreme value distribution，EVD），而不是正态分布。1958 年，Emil Gumbel 在他的《极值统计》一书中描述了 EVD 的统计基础。该图（Gumbel，1958）显示了极值分布，注意对标记有“largest”的曲线，其尾部靠右。同时，如表中所显示，平均值以上三个标准偏差的值以下的部分占据了正态分布曲线下 99.865%的区域，而在 EVD 分布中这个值是 98.810%。换句话说，EVD 分布中图中极值右边区域的面积更大。我们将应用这种分析到 BLAST 搜索结果，以评估一条查询序列是否和数据库中搜索到的结果显著相关。授权使用

第 4 章

局部比对搜索基本工具BLAST

4.1 引言

基本局部比对搜索工具 BLAST 是 NCBI 中用来将一个蛋白质或 DNA 序列和各种数据库中的其他序列进行比对的主要工具（Altschul，1990，1997）。BLAST 搜索是研究一个蛋白质或基因的最基本的方法之一。这种搜索将告诉我们哪些相关的序列在同一物种或其他物种中出现。NCBI 的网页上包含了一些很好的用来学习 BLAST 的资源。

NCBI 上的资源包括可以通过 BLAST 主页页边上的按钮访问的一个指南和一个教程（http://www.ncbi.nlm.nih.gov/BLAST）。

在第 3 章中，我们介绍了怎样进行两个蛋白质或者核酸序列的两两成对比对。BLAST 搜索允许用户选择一个序列（记作查询序列），然后将这个查询序列与一个整个的数据库（记作目标序列）进行两两序列比对。典型的情况下，这意味着在一次 BLAST 搜索中要进行成千上万次的比对分析，然后将最接近的比对结果返回。Needleman-Wunsch（1970）和 Smith-Waterman（1981）全局以及局部比对算法可以找到最优的比对结果，但我们不能用它们来做数据库搜索，因为这些算法是非常消耗计算机资源的（用这些方法将一个查询序列对一个庞大的数据库进行搜索将需要相当多的时间）。就像本章中将要介绍的，BLAST 兼顾了速度和灵敏度。它还提供了方便的互联网访问方式。

BLAST 程序允许用一个 DNA 或者蛋白质查询序列和一个 DNA 或者蛋白质的数据库进行比对。一个 DNA 序列可以转换成 6 个潜在的蛋白质（见下文），而且 BLAST 的算法包含了将蛋白质序列对动态翻译的 DNA 数据库或者反过来进行比对的方法，因此 BLAST 具有非常广泛的应用。这包括以下 7 个方面。

① 确定特定的蛋白质或核酸序列有哪些已知的直系同源或旁系同源序列。除了 RBP 外，还有哪些其他的脂质运载蛋白是我们所知道的？当一个新的细菌的基因组被测序后，几千种蛋白质被确定，其中有多少蛋白质是同源的？从这里面预测出的基因中有多少是在 GenBank 中找不到显著性同源物的？

② 确定哪些蛋白质和基因在特定的物种中出现。植物中是否也存在像 RBP 这样的脂质运载蛋白？鱼类中是否有反转录酶基因（如 HIV-1 *pol* 基因）？

③ 确定一个 DNA 或蛋白质序列身份。如你可能通过一个负杂交实验或芯片实验发现一个特殊

的 DNA 序列在你所使用的实验条件下是被显著调控的，那么就可以通过将这个 DNA 序列在一个蛋白质数据库中进行搜索，来寻找哪些蛋白质是与你的 DNA 序列所编码的蛋白质相关性最高的。

④ 发现新基因。例如，一个对于全基因组 DNA 的 BLAST 搜索可能会发现一个 DNA 所编码的蛋白质是以前所没有报道过的。在第 5 章里，我们将向你展示怎样用 BLAST 搜索来发现新的、以前没有研究过的基因。

⑤ 确定一个特定基因或者蛋白质有哪些已经被发现了的变种。例如，很多病毒都具有极强的突变能力；HIV-1 pol 有哪些已知的变异体？

⑥ 研究可能存在多种剪接方式的表达序列标签。有专门用于 BLAST 搜索的 EST 数据库。实际上有许多用来进行搜索的专门的数据库。例如，专门的包含同一个特定的物种、一种组织、一个染色体、一种 DNA（如非转录区），或者一个蛋白质功能类的序列数据库。

⑦ 寻找对于一个蛋白质的功能和/或结构起关键作用的氨基酸残基。一次 BLAST 搜索的结果可以被放在一起比对（见第 10 章），这时候，我们就会发现其中那些像半胱氨酸残基这样可能具有重要生物学功能的保守残基。

BLAST 是包含一组可以输入查询序列，然后将其和一个数据库中的 DNA 或蛋白质序列进行比较的程序。这些程序会产生表示你的查询序列和数据库中序列进行局部比对结果的高分片段对（HSPs）。

进行任何的 BLAST 搜索都包括以下 4 个步骤。

① 选择你所感兴趣的序列，将它粘贴到 BLAST 的输入框中。

② 选择一个 BLAST 程序（blastp，blastn，blastx，tblastx，tblastn）。

③ 选择一个用于搜索的数据库。一个通常的选择是去冗余（nr）数据库，但还有很多其他数据库。

④ 为搜索和输出格式选择可选参数。这些选项包括选择替换矩阵，过滤复杂度低的序列，以及将搜索范围限制在某些特定的物种中。

可以通过 http://www.ncbi.nlm.nih.gov/BLAST 直接访问 BLAST 的网站，或者访问 NCBI 的主页（http://www.ncbi.nlm.nih.gov），然后点击工具按钮中的 BLAST 按钮。

当我们介绍 BLAST 搜索的步骤时，我们可以从一个特殊的例子开始。选择链接 “Standard protein-protein BLAST [blastp].” 你将看到一个标有 “search” 的输入框，输入人类 RBP4 的访问编号（NP _ 006735），然后点击 “BLAST!” 按钮（图 4-1）。会打开一

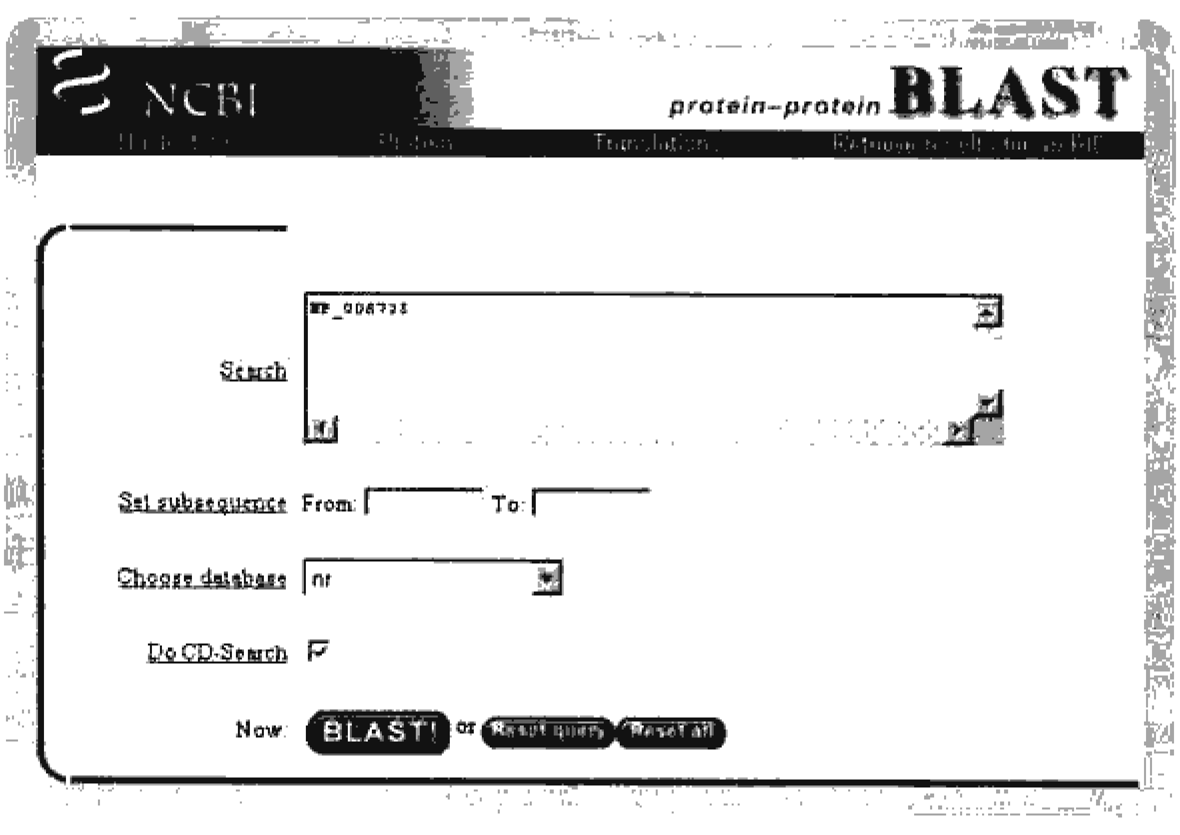


图 4-1　对人类的 RBP4 蛋白进行蛋白-蛋白 BLAST

个名为“Formating BLAST”的新页面（见图 4-2）。点击“format”按钮（图 4-2）等待你的结果。结果中列出了与 RB 关系最密切的蛋白。下面我们将具体介绍 BLAST 搜索的一些实际情况。

按照 2003 年 5 月的统计，你已经在几秒钟内搜索了一个包含超过 150 万个蛋白质序列（和大约 5 亿个氨基酸残基）的数据库。

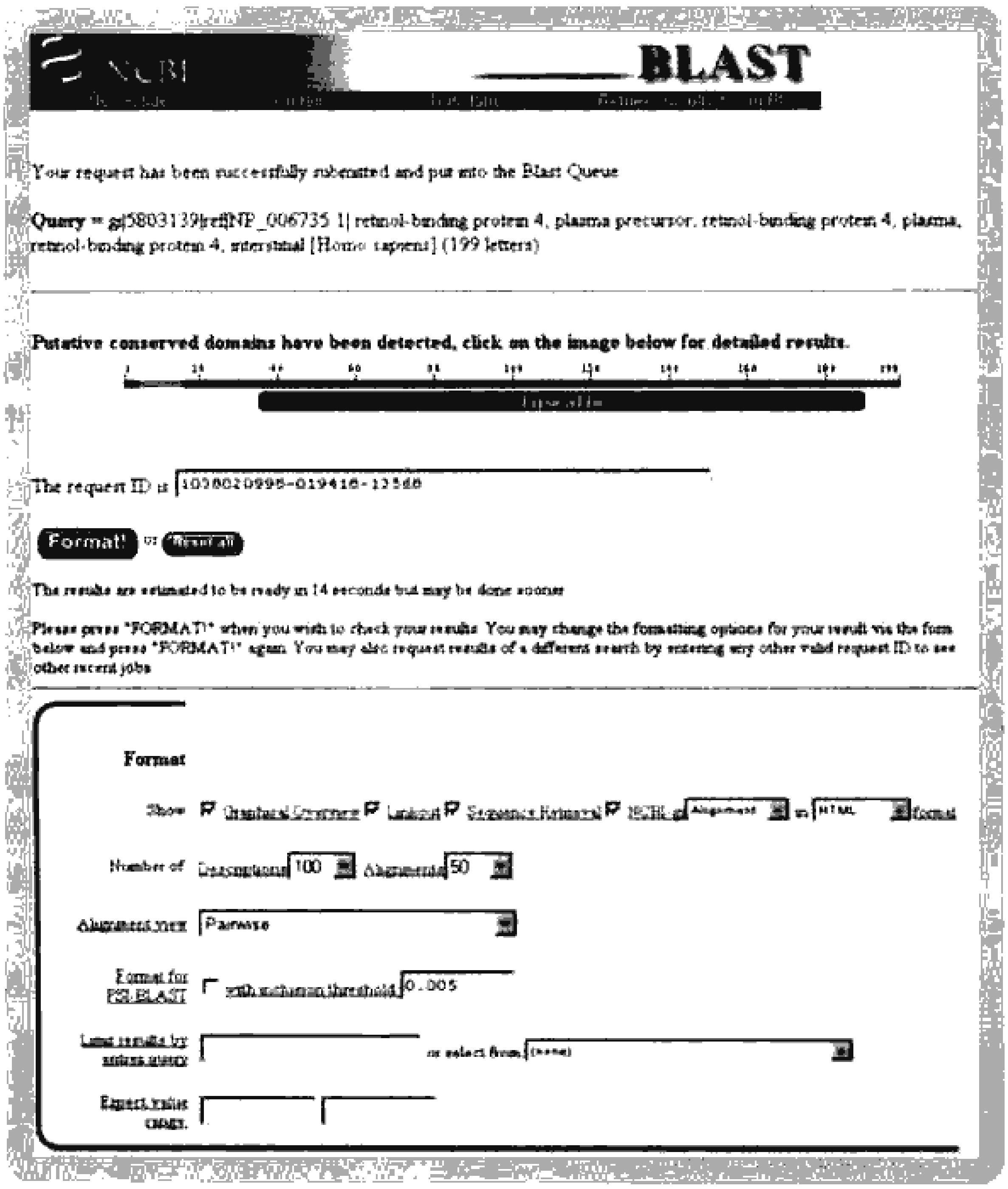


图 4-2 Formating 的 BLAST 页

4.2 BLAST 搜索步骤

4.2.1 步骤 1：选定感兴趣的序列

要进行一次 BLAST 搜索首先要选择一个 DNA 或者蛋白质序列。一共有 3 种主要的输入方式：①剪切然后粘贴 DNA 或者蛋白质序列；②使用 FASTA 格式的序列；③简单地使用访问编号［例如，一个 RefSeq 或 GenBank（GI）的序号］。一个 FASTA 格式的序列以一个单行的说明开始，接下来是若干行的序列数据。说明行以其开始处的一个大于号（>）与序列数据区分开来。我们建议所有的文本行的长度都少于 80 个字符。图 2-10 是一个 FASTA 格式序列的例子。

有关 FASTA 格式在 http://www.ncbi.nlm.nih.gov/BLAST/fasta.html 上有更为详细的介绍。注意不要把 FASTA 格式与我们在第 3 章中所介绍的 FASTA 程序弄混了。

在一个 BLAST 搜索中输入访问编号通常要简单和容易一些。注意 BLAST 程序可以识别和忽略出现在你的输入序列字母中间的数字。BLAST 搜索还允许你输入一个完整序列的一个子集，例如你感兴趣的一个区域或结构域。

对于 BLAST 搜索，你的查询序列可以是大写也可以是小写，中间可以出现也可以不出现空格或数字（对于 FASTA 序列，格式则是不同的，中间是不能出现空格的）。如果查询的是 DNA，BLAST 算法将会对两条链都做搜索。

4.2.2　步骤 2：选择 BLAST 程序

NCBI 的 BLAST 程序家族包括 5 个主要的程序。总结见图 4-3。

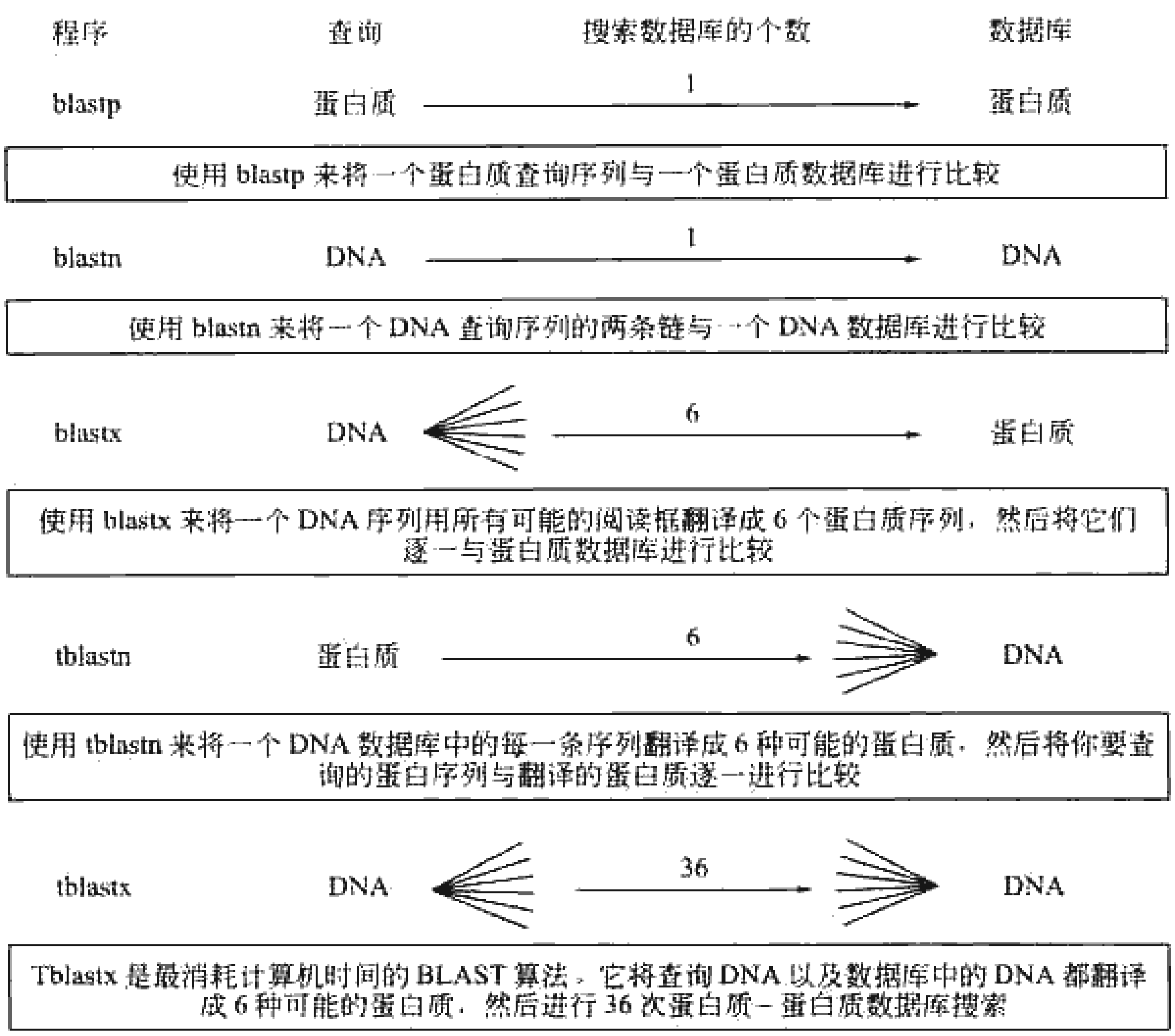


图 4-3　5 个 BLAST 算法程序简图。后缀 p 表示蛋白质（如在 blastp 中）；n 表示核苷酸，x 表示一个 DNA 查询被动态的翻译成 6 种蛋白质序列。前缀 t 表示"翻译（translating）"，也就是将一个 DNA 数据库动态地翻译成 6 种蛋白质。我们将在本章的后面部分中讨论这些 BLAST 方法的用途

① 第一个程序是 blastp。这个程序将一个氨基酸的查询序列与一个蛋白质序列数据库相比较。注意这类搜索是有专门与蛋白搜索相关的可选参数（见下文）的，例如对各种 PAM 和 BLOSUM 打分矩阵的选择。

② 第二个程序是 blastn。这是用来将一个核酸的查询序列与一个核酸序列数据库相比较的。

另外三种 BLAST 算法是依赖于 DNA 与蛋白质间的基本关系的。任何 DNA 序列都可以被转录并翻译成 6 种可能的阅读框（三种在上面的链上，三种在下面的链上，如图 4-4）。在 BLAST 搜索中，查询的 DNA 序列可能被翻译成可能的蛋白质，整个一个 DNA 数据库也可能被翻译，或者以上两者都被翻译。在所有这三种情况下，这些算法将进行蛋白质-蛋

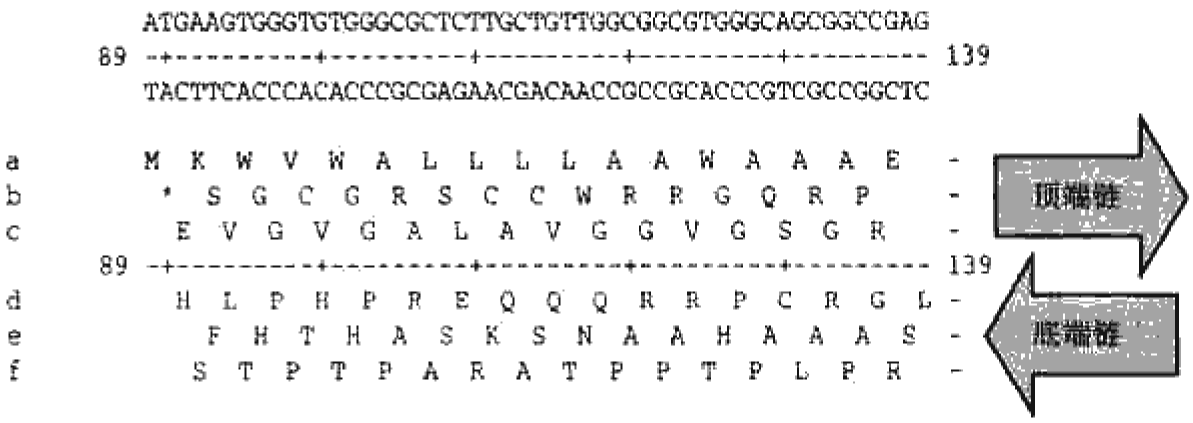


图 4-4 DNA 可能编码 6 种不同的蛋白质。图中是人类 RBP（访问编号：NM _ 006744）的两条链的 DNA 序列。在顶端链上可以编码 3 种可能的蛋白质（阅读框 a、阅读框 b、阅读框 c）。阅读框 a 编码的是真正的 RBP 蛋白。阅读框 b 的第一个密码子 TGA 是一个终止密码子（图中的星号）。在底端链上，可能编码另外 3 种蛋白质（阅读框 d、阅读框 e、阅读框 f）。例如，蛋白质 d 的第一个氨基酸是亮氨酸，由密码子 CTC 编码。该图是用作图程序 GCG 制作的

白质比对。

③ blastx 程序将一个核酸的查询序列按所有可能的阅读框翻译后的序列与一个蛋白质序列数据库进行比较。如果你有一个 DNA 序列，想知道它编码什么蛋白质（如果它编码蛋白质的话），你就可以使用 blastx 搜索。它会自动地将 DNA 翻译成 6 种可能的蛋白质（图 4-3 和图 4-4）。然后 blastx 程序就会将翻译的 6 个蛋白质序列逐一地与蛋白质数据库中的每个成员进行比较。

④ tblastn 程序将一个蛋白质查询序列与一个以所有阅读框动态翻译成蛋白的核酸序列数据库进行比较。你可以使用这个程序来判断一个 DNA 数据库是否编码所感兴趣的查询蛋白。用 RBP 来查询是否可以在某个已测序的 DNA 数据库中找到匹配项呢？

⑤ tblastx 程序将一个核酸查询序列的 6 种框架的翻译结果与一个核酸序列数据库的 6 种框架翻译产物进行比较。tblastx 程序不能使用 BLAST 网页上提供的主要的去冗余（nr）数据库，因为这一操作是非常消耗计算机资源的。我们将在下面给出去冗余数据库的定义。假设你有一个 DNA 序列在数据库中找不到明显的匹配，现在你想知道它是否编码一个与数据库中的成员所编码蛋白也具有非常远的亲缘关系的蛋白质。这时 blastx 搜索将会是非常有用的，它将帮你找出那些有亲缘关系的项目。但如果这个搜索失败了，你还可以进行一次 tblastx 搜索来确定整个 DNA 数据库中是否包含编码与你的查询序列编码蛋白同源的蛋白质的基因。

4.2.3 步骤 3：选择数据库

BLAST 搜索可以使用的数据库会列在每一个 BLAST 页面上，对于蛋白质数据库搜索（blastp 和 blastx），两个主要的选择就是 nr 数据库和 SwissProt。对于蛋白质来讲，nr 数据库包含联合了 GenBank、蛋白质数据银行（PDB）、SwissProt、PIR 和 PRF（参见第 2 章关于这些数据库的介绍）的蛋白质记录。

> 有关数据库的信息还可以从 http://www.ncbi.nlm.nih.gov/blast/html/blastcgihelp.html 中获得。

对于 DNA 数据库搜索（blastn、tblastn、tblastx）可供选择的有核苷酸 nr 数据库或者 EST 数据库。用于 DNA 搜索的 nr 数据

库包括 GenBank、EMBL、DDBJ，以及 PDB 中的核苷酸序列。但是，nr 数据库中不包括 EST、STS、GSS 或者高通量基因组测序（HTGS）数据库中的记录。

nr 数据库是合并了若干个主要的蛋白质或 DNA 数据库得到的。这些数据库中经常包含有相同的序列，但 nr 数据库只收录其中的一个序列（即使在 nr 数据库中出现看上去一样的序列，它们实际上还是具有一些细节上的区别的）。nr 数据库是在要搜索现有的绝大多数序列时典型和常用的数据库。

表 4-1 和表 4-2 提供了对于 NCBI 上的标准 BLAST 搜索程序可以使用的所有蛋白质序列和核酸序列数据库的总结。

表 4-1 标准 BLAST 搜索可以使用的蛋白质序列数据库

数据库	描述
nr	去冗余 GenBank 编码序列＋PDB＋SwissProt＋PIR＋PRF
Month	最近 30d 内发布的序列数据
SwissProt	SwissProt 最新发布的数据
Drosophila	果蝇基因组计划得到的果蝇蛋白质(http://www.fruitfly.org)
S. cerevisiae	酵母(*Saccharomyces cerevisiae*)蛋白质
E. coli	大肠杆菌(*Escherichia coli*)蛋白质
Pdb	Brookhaven 的蛋白质数据库(http://www.rcsb.org/pdb/)
alu	翻译后的从 REPBASE 中选出的 Alu 重复序列，适用于遮盖查询序列中的 Alu 重复，可以通过匿名文件传输协议(FTP)从 ncbi.nlm.nih.gov(在/pub/jmc/alu 目录下)得到

注：改写自 http://www.ncbi.nlm.nih.gov/blast/html/blastcgihelp.html#protein.databases。

表 4-2 标准 BLAST 搜索可以使用的核酸序列数据库

数据库	描述
nr	所有 GenBank＋RefSeq＋Nucleotides＋EMBL＋DDBJ＋PDB 数据库中的序列(但不包括 EST，STS，GSS，或 0 阶段，1 阶段和 2 阶段的 HTGS 序列)。不再是"去冗余"的了
month	GenBank＋EMBL＋DDBJ＋PDB 数据库中最近 30d 内发布的所有新的或者是修改了的序列
Drosophila genome	由 Celera 公司和 Berkeley 果蝇基因组计划(BDGP)所提供的果蝇(Drosophila)基因组(http://www.fruitfly.org/)
Dbest	由 GenBank＋EMBL＋DDBJ 数据库中数据按 EST 划分形成的数据库
Dbsts	由 GenBank＋EMBL＋DDBJ 数据库中数据按 STS 划分形成的数据库
Htgs	为完成的高通量基因组序列数据
Gss	基因组抽样序列，包括单次基因组数据、外显子捕捉序列和 *Alu* PCR 序列
S. cerevisiae	酵母(*Saccharomyces cerevisiae*)基因组核酸序列
E. coli	大肠杆菌(*Escherichia coli*)基因组核酸序列
Pdb	从 Brookhaven 的蛋白质数据库的三维结构中得到的序列数据(http://www.rcsb.org/pdb/)
Vector	GenBank、NCBI 中的载体的子集，在 ftp://ncbi.nlm.nih.gov/blast/db/上
Mito	线粒体序列的数据库
alu	从 REPBASE 中选出的 Alu 重复序列，适用于遮盖查询序列中的 Alu 重复(可以通过匿名 FTP ncbi.nlm.nih.gov 在/pub/jmc/alu 目录下得到)
epd	真核生物启动子数据库(http://www.genome.ad.jp/dbget-bin/www_bfind? epd)

注：改写自 http://www.ncbi.nlm.nih.gov/blast/html/blastcgihelp.html#nucleotide.databases。

4.2.4 步骤 4：选择参数

4.2.4.1 步骤 4a：选择可选的搜索参数

一开始先把注意力放在一个标准的蛋白质-蛋白质 BLAST 搜索上。当确定了要输入的序列和要搜索的数据库之后，还有 10 个其他的可选参数等着去确定（图 4-1 和图 4-5）。

保守结构域数据库位于 http://www.ncbi.nlm.gov/Structure/cdd/cdd.shtml。

① Do CD-Search（图 4-1）。该选项是关于保守结构域数据库（CDD）的，它可以在一个蛋白质中找到保守的模块或结构域。我们将在第 8 章中具体介绍 CDD 数据库。

② Limit by Entrez Query（图 4-5）。任何 NCBI BLAST 搜索的范围都可以用在 Entrez 搜索中使用的任何一种范围限定词来限定。例如在该选项中键入限定词"protease"，使用查询蛋白 RBP4（NP _ 006735）进行 blastp 搜索：则得到的比对结果不再是很多条，而是得到一个与具有丝氨酸蛋白酶抑制剂结构域的蛋白质融合的脂质运载蛋白（α-1-微球蛋白；NP _ 001624）。BLAST 搜索也可以通过一个方便的下拉菜单来限制将要搜索的物种。一些常用的物种包括古细菌、后生动物（多细胞动物）、细菌、脊椎动物、真核生物、哺乳动物、Embryophtya（高等植物）、啮齿动物、真菌和灵长类动物。BLAST 搜索的范围可以被限制在任何一个属和种或者其他分类学上的类中。

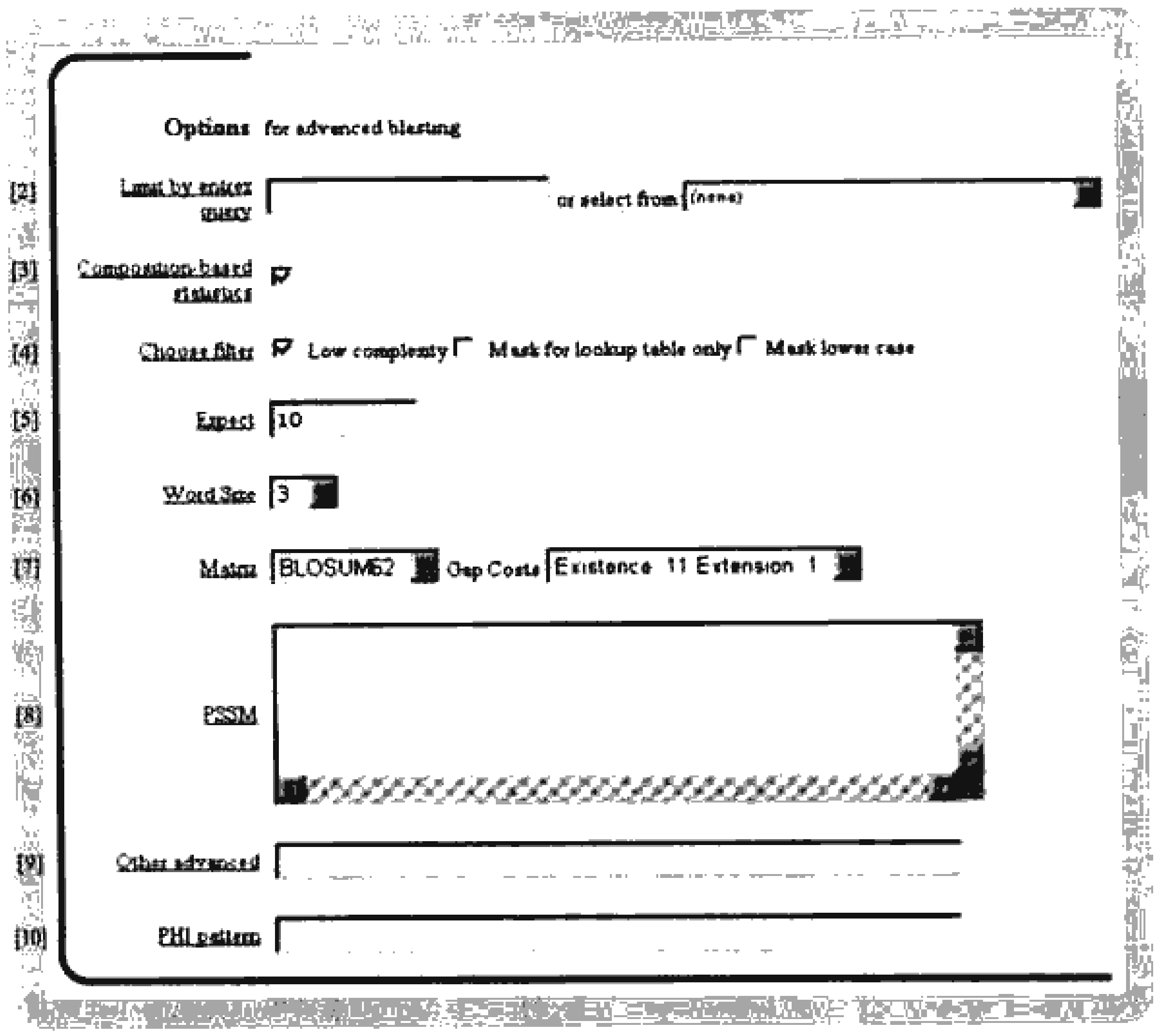


图 4-5 一个蛋白质-蛋白质 BLAST 页面中的选项菜单（输入框中的数字参见下文的讨论）

③ Composition-Based Statistics。这个选项是默认选择的，一般来说可以改善 E 值的统计计算（参见下文）和提高灵敏度（减少返回的假阳性结果的数目）。这个选项对于每一个数据库序列使用一个略有不同的打分系统，对于 PSI-BLAST 的位置特异性打分矩阵尤其具有重要的应用性。

④ 选择过滤条件。过滤掉查询序列中具有较低复杂度的掩盖部分（或者具有很大偏差的成分）（Wootton 和 Federhen，1996）。低复杂度序列即指那些非常常见的含有非常有限信息量的氨基酸（或者核苷酸）序列段。这样的例子包括双核苷酸重复序列，或者蛋白质序列中极端富含某一种或某两种氨基酸的区域。由疏水性氨基酸残基组成的用来形成跨膜结构域的序列段是很常见的，因此一个用这样的序列进行的数据库搜索将会得到大量具有统计上的显著性，但几乎没有任何生物学上的关联意义的匹配结果。其他一些在过滤过程中被掩盖

的模式段包括富含酸性残基、碱性残基和脯氨酸残基的区域。对于蛋白质查询序列，使用 SEG 程序进行过滤；对于核酸查询序列则使用 DUST 程序。对于一个经过了过滤处理的查询序列，邻近的低复杂度残基被换成了一串字母 X（对于蛋白质序列；见图 4-6）或者字母 N（对于核酸序列）。当使用一个很短的 DNA 或者蛋白质序列来查询时，则关掉过滤选项将很可能会很有帮助。这里要注意的是被过滤的是所提交的查询序列而不是数据库。

```
>gi|5803139|ref|NP_006735.1| retinol-binding protein 4, interstitial precursor [Homo sapiens]
 gi|132404|sp|P02753|RETB_HUMAN PLASMA RETINOL-BINDING PROTEIN PRECURSOR (PRBP) (RBP)
 gi|72065|pir||VAHU plasma retinol-binding protein precursor - human
 gi|35897|emb|CAA24959.1| (X00129) precursor RBP [Homo sapiens]
          Length = 199

 Score =  378 bits (971), Expect = e-104
 Identities = 187/199 (93%), Positives = 187/199 (93%)

Query: 1    MKWVXXXXXXXXXXXXERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAE 60
             MKWV            ERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAE
Sbjct: 1    MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAE 60

Query: 61   FSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 120
             FSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH
Sbjct: 61   FSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 120

Query: 121  WIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQ 180
             WIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQ
Sbjct: 121  WIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQ 180

Query: 181  YRLIVHNGYCDGRSERNLL 199
             YRLIVHNGYCDGRSERNLL
Sbjct: 181  YRLIVHNGYCDGRSERNLL 199
```

图 4-6　一个选择了过滤选项的 BLAST 搜索的输出结果。注意查询序列中的 12 个氨基酸已经被字母 X 所代替。这些残基因为低的复杂度而被过滤掉了。在本例中，这 12 个氨基酸是信号序列的一部分并且是疏水的；如果不进行过滤，成百的同样具有疏水性氨基酸片段的蛋白质作为假的数据库匹配也会出现在结果中。对于核苷酸 BLAST 查询，低复杂度序列（如单核苷酸或双核苷酸串）将会被字母 N 取代

过滤选项的调整对于 BLAST 搜索的结果将会产生显著的影响。当使用 blastp nr 来查询一个人类富含脯氨酸的蛋白质（NP _ 036522）时，会得到 8 个数据库中的匹配项，而其中两个具有统计显著性（图 4-7）。而当关掉过滤选项后再搜索，就会得到包含其他一些富含脯氨酸的蛋白质在内的 23 个匹配（图 4-8）。一般来讲，过滤可以帮助避免那些假的数据库匹配，但在某些情况下一些可信的匹配也可能会被过滤掉。

⑤ 期望。期望值 E 是得分大于或等于某个分值 S 的不同的比对的数目在随机的数据库搜索中发生的可能性。如图 4-7 中所示的那些最佳匹配，其得分是 45 比特，E 值为 1e-04（也就是 1×10^{-4}，或 0.0001）。这表明在使用这些特定搜索参数（包括数据库的规模以及打分矩阵的选择）的情况下，打 45 分或者更好的情况发生的概率是 1/10000。

期望值（expect value）有时候也写成 expectation value。

对于 blastn、blastp、blastx 和 tblastn 期望值的默认设定值是 10。在这个 E 值下，随机出现得分等于或高于比对得分 S 的期望数为 10 个（这里是假设用与实际的查询序列长度相等的随机的查询序列搜索数据库）。当将期望选项值调小时，返回的数据库搜索结果将变少；匹配被搜索到的概率也会变小。增大 E 值将返回更多的结果。我们将在本章的稍后部分讨论 BLAST 搜索的统计学时更详细地介绍 E 值。

⑥ 字段长度。对于蛋白质搜索，窗口大小可以被设定为 3（默认值）或者 2。当用一个查询序列来进行数据库搜索时，BLAST 算法首先将查询序列分割成一系列具有特定长度（字段长度）的小的序列段（字段），我们将在下文中进行详细的描述。对于 blastp，更大的

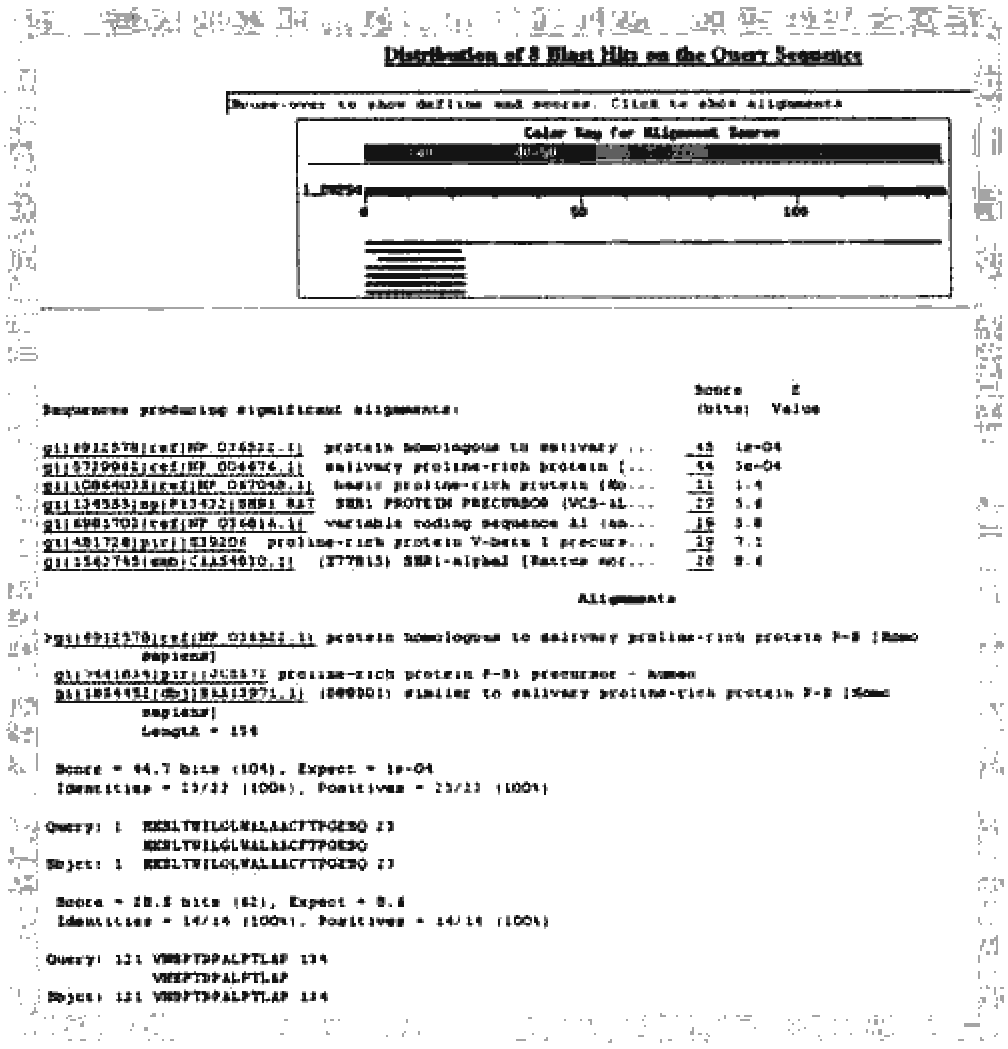


图 4-7 以一种富含脯氨酸的人类唾液蛋白（NP_036522）作为查询项使用 blastp 对于 nr 数据库进行搜索所得到的结果。这里只得到了 8 个数据库匹配项，其中两个与查询项具有较近的亲缘关系。同样，该蛋白质的大部分（氨基酸 24～120）没有在数据库中的匹配，这是由于该蛋白的中间区域被默认的过滤功能给过滤掉了

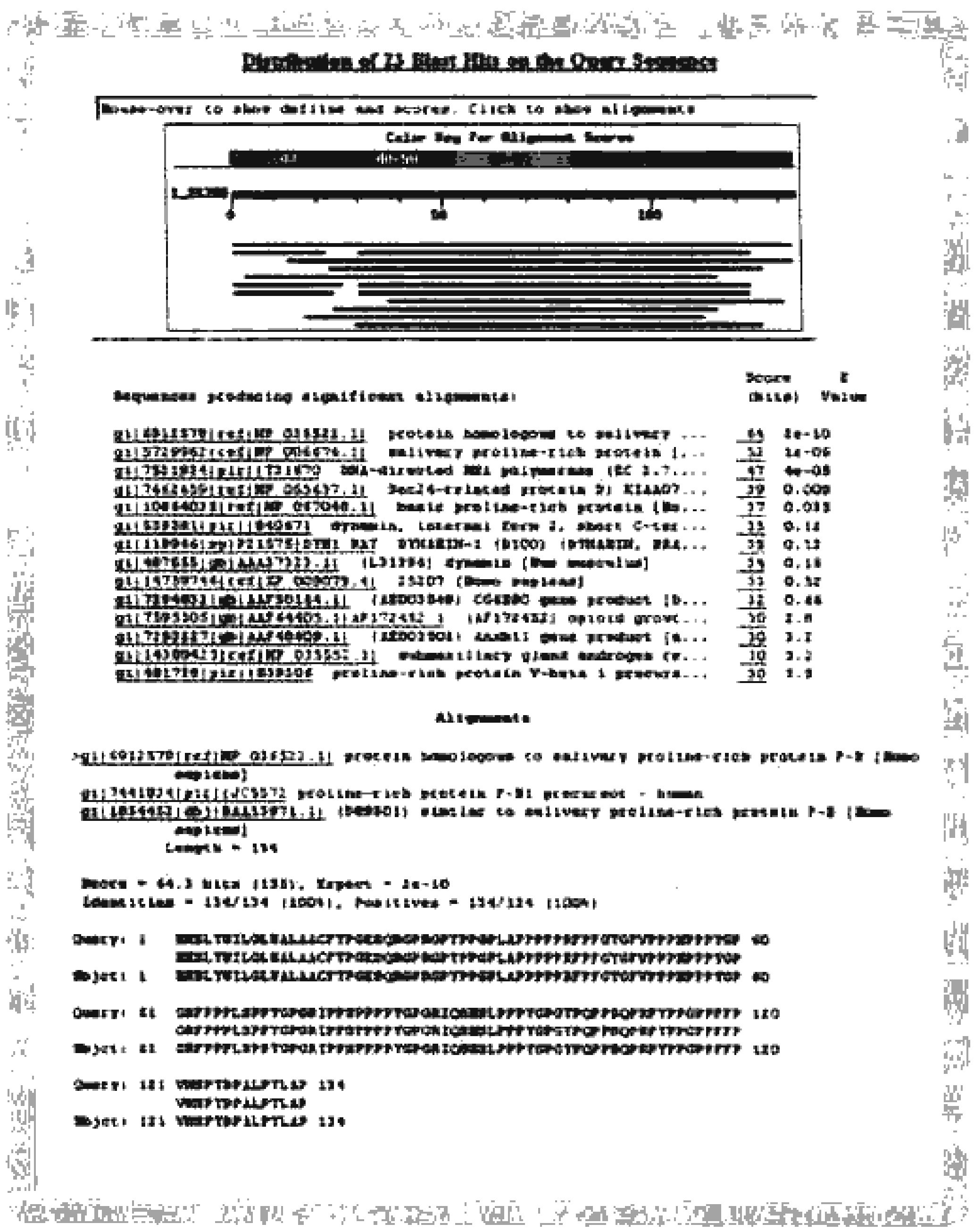


图 4-8 关掉过滤选项使用一个富含脯氨酸的蛋白质进行的 blastp 搜索结果。这次得到了更多的与在蛋白质的全序列范围内相匹配的数据库返回结果（与图 4-7 相比较）。第一个比对显示了该蛋白极度富含脯氨酸的特点，而这个特点会被 BLAST 的默认过滤功能过滤掉

尝试用 RBP 蛋白（NP_006735）作 blastp 搜索，分别选用 E 值为 1000、10（默认值）或者 0.0001。在每种情况下，无论他们显著与否，BLAST 算法都得到相同数目的数据库匹配。但是 blastp 程序在它的输出结果中给出的列表中包含大约 750 个 E 值比 1000 好的序列、140 个 E 值比 10 好的序列、50 个 E 值比 0.0001 好的序列。我们将在图 4-16 时看到如何在 BLAST 搜索输出页的最底部找到这些信息。

字段长度将得到更高的搜索精度。对于任意的字段长度，每个字段的匹配结果将被延伸以得到 BLAST 的输出结果。实际应用中对于蛋白质搜索很少需要改变字段的长度。

对于核酸搜索，默认的字段长度是 11，它可以被增大（字段长度为 15）或者减小（字段长度为 7）。降低字段长度将会使搜索变得更准确同时也会变得更慢。

⑦ 矩阵。对于 blastp 的蛋白质-蛋白质搜索有 5 种氨基酸替代矩阵：PAM30，PAM70，BLOSUM45，BLOSUM62（默认值）以及 BLOSUM80。一些其他的 BLAST 服务器（参见第 5 章中关于高级 BLAST 搜索的部分）还提供了很多其他的替代矩阵，如 PAM250。通常情况下明智的选择是在一次 BLAST 搜索中使用几种不同的打分矩阵。例如，就像在第 3 章中所提到的，PAM40 和 PAM250（参见图 3-16）作为打分矩阵具有完全不同的特性，可以反映序列在不同层次上具有相似性。

⑧～⑩ 关于位置特异性打分矩阵（PSSM）、其他一些高级选项，以及模式匹配启动矩阵（PHI）将在第 5 章中进行讨论。

4.2.4.2　步骤 4b：选择可选格式参数

BLAST 搜索有很多控制输出格式的参数。这里通过一个用人类 RBP4 蛋白（NP_006735）作为查询，并将搜索的物种范围限制为老鼠（*Rattus norvegicus*）的蛋白质-蛋白质 BLAST 搜索的例子来说明。这个搜索的结果分成几个主要部分。在顶部（图 4-9）提供了关于该搜索的详细信息，包括 BLAST 搜索的类型、关于查询内容和所搜索的数据库的描述，以及一个分类连接可以将结果按照物种进行分类。

一个典型的 BLAST 搜索结果的中间部分显示的是数据库中序列与查询序列相匹配的项的列表（图 4-10）。简明图形提供了用不同颜色表示的搜索结果的概况。图下面的每一个条带表示数据库中的一个与查询序列相匹配的蛋白质或核酸序列。每一个条带的位置都与查询序列

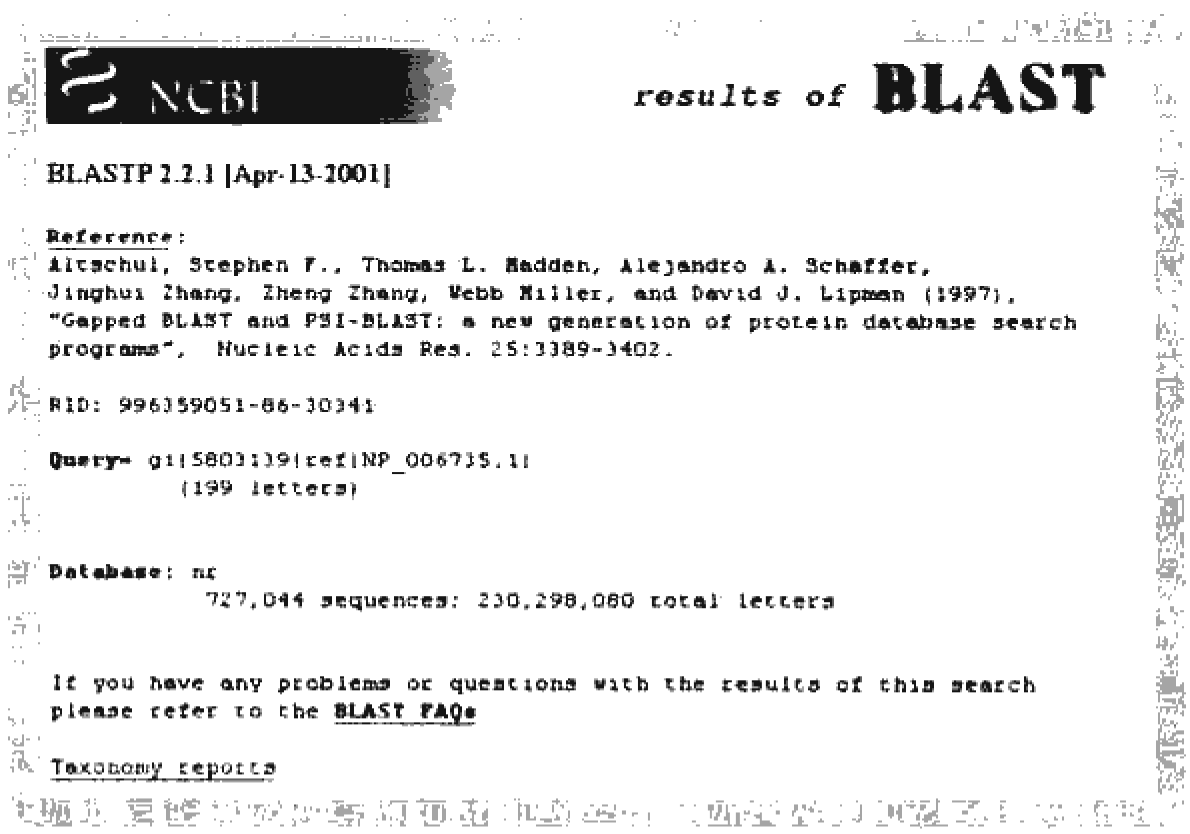


图 4-9　BLAST 输出的顶端部分描述了所进行的搜索（在本例中为 BLASTP 2.2.1）、查询序列（人类 RBP）、所搜索的数据库（nr 还包含了该数据库的大小）。这里还提供了一个得到将搜索结果按物种分类后的分类报告的链接

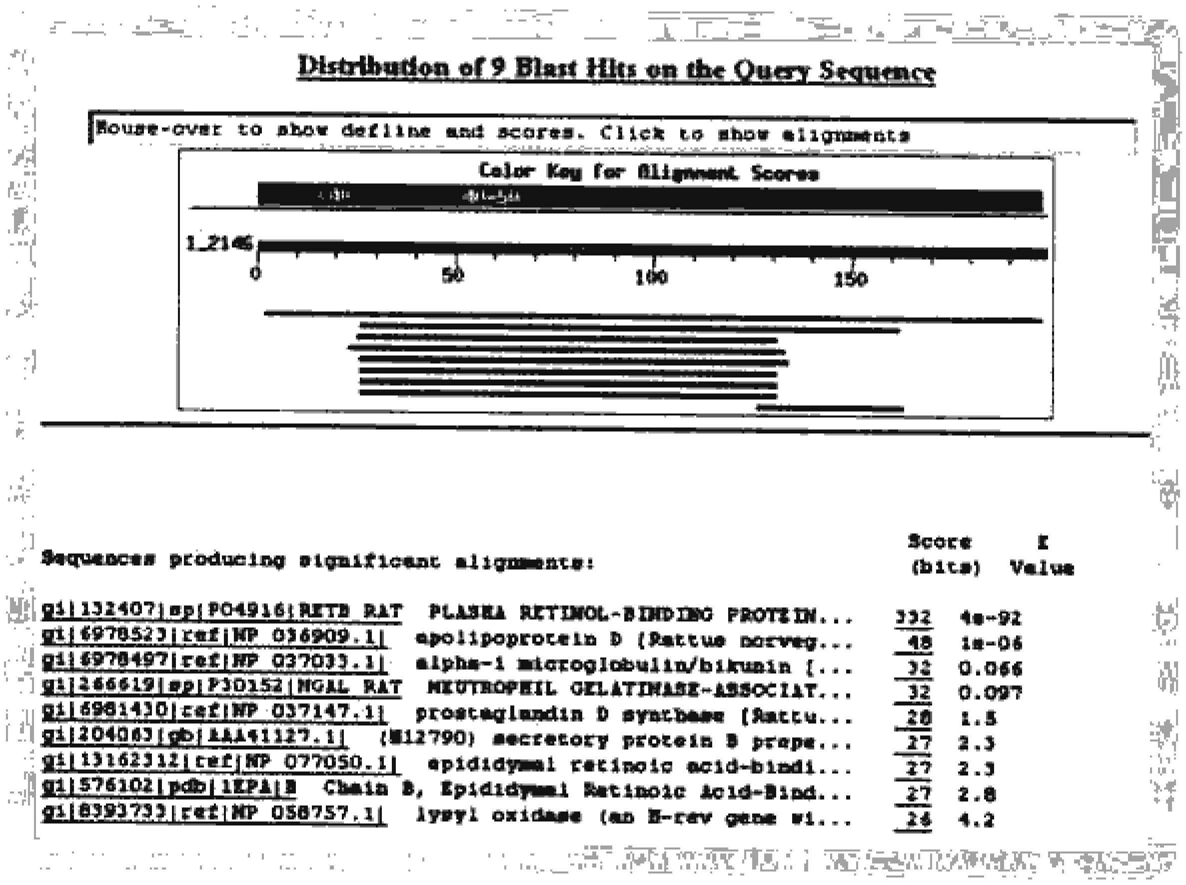


图 4-10 一个典型的BLAST搜索结果的中间部分提供了结果的图示输出，然后是那些数据库序列的列表。在图示结果中，数据库匹配被标以不同颜色以表示亲缘关系的远近（根据比对得分），每条线的长度对应于该序列与查询序列比对上的区域大小。与查询序列比对的序列的列表中包含有到数据库中记录和到相应比对的链接。同时还提供了每一个比对的比特分数和 E 值。注意位于列表顶端的最佳匹配具有大的比特分数和小的 E 值

的线状图的相应部位相对应，从而使用户立即看到数据库匹配与查询序列比对上的某一个或多个区域的位置。最接近的匹配将出现在图的最上部并用红色表示。阴影区域（如果出现的话）对应于在同一个数据库查询中所找到的两个或更多具有相似性的不同区域中的序列差异的部分。

所有比对用一个被称为“描述”的单行小结的列表来描述。这些描述行按照 E 值增加的顺序排序；因此显著性最强的（E 值最低的）匹配将位于最顶端。描述包括以下4个项目（按照从左向右的顺序）。

① 数据库序列的标号。下面是一些常见的序列标号：

RefSeq(e. g. ,ref|NP_006735. 1|)；

Protein database(e. g. ,pdb|1RBP|)；

Protein Information Resource(pir||A27786)；

Swiss-Prot database(sp|P27485|)；

GenBank(e. g. ,gb|AAF69622. 1|)；

European Molecular Biology Lab(EMBL)(e. g. ,emb|CAB64947. 1|)，以及 Database of Japan(DBJ)(dbj|BAA13453. 1)。

② 关于序列的简单描述。

③ 对于每一个数据库序列找到的得分最高的匹配的比特分数（比特分数的概念将在下文给出）。

④ 期望值 E。标号可以链接到GenBank中完整的记录。点击描述行中给出的分数将把读者带到相应的序列比对。也可以通过向下滚动输出页看到比对。

BLAST搜索输出结果的靠下面部分是一系列的两两序列比对，就像图4-6中的例子那样。在这里可以检查查询序列（输入序列）与对象序列（如和查询序列比对的特定的数据库匹配）之间的比对情况。这里提供了4种衡量的分数：比特分数、期望分数、一致性百分比、正性（相似性百分比）。

可以不用整体地重新进行BLAST搜索，而通过改变格式选项来提供一些不同的输出选择（图4-11）。描述以及比对的数目都是可以改变的。一些选项还可以把比对序列显示成对

序列比对的形式（图 4-12），这对于确定在一个蛋白质或 DNA 家族中的保守的（或趋异的）氨基酸残基是非常有用的。

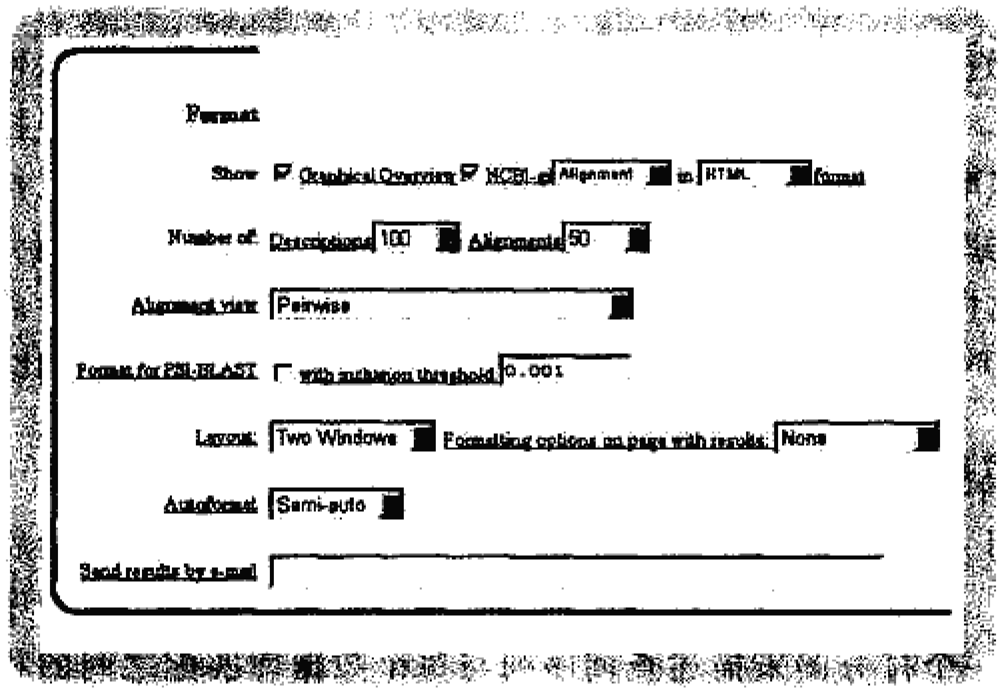


图 4-11　一个蛋白质-蛋白质搜索的格式选项

```
□ 1_24439    3   VVXXXXXXXXXXXXXERDCRVSSFRVKENFDKARFSGTWYANA--------------KKDPE 49
□ 27685009   5   WALVLLAALGGGSAERDCRVSSFRVKENFDKARFSGLWYAIA--------------KKDPE 51
□ 6978523    33                             VQENFDVKKYLGRWYEI---------------EKIPV 54
□ 6978497    29                            QVQENFNEARIYGKWFNLAVGSTCPWLRRIKNKMSVS 65
□ 18543345   35                          SVPLQPGFWTERFQGRWFVVGLAANAV--------QKERQ 66
□ 27706210   25                             VGKNYNMARISGVWHLVS------------------ 42
□ 6981430    30                             VQPNFQQDKFLGRWYSAGLASNSSWF-----REKKE 60
□ 27684523   202                                    SGFW---A--------------KALPE 211
□ 204063     20                             VVKDFDISKFLGFWYEIAFAS-----------KMGTP 45
□ 13162312   24                             VVKDFDISKFLGFWYEIAFAS-----------KMGTP 49
□ 576101     2                              VVKDFDISKFLGFWYEIAFAS-----------KMGTP 27

□ 1_24439    50  GLFLQDNIV-AEFSVDETGQMSATAKGRVR---LLN----NW---DV------------CA 87
□ 27685009   52  GLFLQDNII-AEFSVDEKGHMSATAKGRVR---LLS----NW---EV------------CA 89
□ 6978523    55  SFEKGNCIQ-ANYSLMENGNIKVLNK-ELRPDGTLN----QV---EG------------EA 94
□ 6978497    66  TLVLQEGATEAEISVTST-----------------------QWRK-GV------------CE 91
□ 18543345   67  SRFTMYSTI-YELQEDNSYNVTS-----------ILV----RG---QG------------CR 97
□ 27706210   43  --MASDNHT----YIEEKGDLRLFIRN-IQ---FLN----NS---NLQFDFHINIHGECV 85
□ 6981430    61  LLFHCQTVV-A------------PSTEGGLN---LTS----TFLRKNQ------------CE 90
□ 27684523   212 EFANEGNII-A-FWVDK---------KGRVF---YRI----N----ES-------------AA 239
□ 204063     46  GLAHKE----------EKMGAMVVELKENLL---ALTTTYYSE---DH------------CV 79
□ 13162312   50  GLAHKE----------EKMGAMVVELKENLL---ALTTTYYSE---DH------------CV 83
□ 576101     28  GLAHKE----------EKMGAMVVELKENLL---ALTTTYYSE---DH------------CV 61

□ 1_24439    88  DMVGTFTDTEDPAKF---KMK--YWGVASFLQ-----KGNDDH-----WIVDTDYDTYAV 132
□ 27685009   90  DMVGTFTDTEDPAKF---KMK--YWGVASFLQ-----RGNDDH-----WIIDTDYDTFAL 134
□ 6978523    95  KQ----SNMSEPAKL---EVQ--FFSLMP-----------PAPY-----WILATDYESYAL 130
□ 6978497    92  EISGVYQKTDIDGKFLYHKSK--W---------------NATLES-----YVVHTNYDEYAI 131
□ 18543345   98  YWIRTFVPSSRPGQY---TLG----NIHSYPQ-----IQSYDV-----QVADTDYDQFAM 140
□ 27706210   86  AVTVVCEKTKMSGEF---SIA--Y---------------EGENKV-----LLVETDYTHYAI 122
□ 6981430    91  TKVMVLQPAGVPGQY---TYNSPHWGSFHSLS-----------------VVETDYDEYAF 130
□ 27684523   240 HLYFSGVRTADPLWA---LVD--VYGLTRGVQLLVAGAGTDLHSHESWWLLVAQEQLYGF 294
□ 204063     80  LEKVTATEGDGPAKF---QVT--RL--------------SGKKEV-----VVEATDYLTYAI 117
□ 13162312   84  LEKVTATEGDGPAKF---QVT--RL--------------SGKKEV-----VVEATDYLTYAI 121
□ 576101     62  LEKVTATEGDGPAKF---QVT--RL--------------SGKKEV-----VVEATDYLTYAI 99

□ 1_24439    133 QYSCR----LLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNG 188
□ 27685009   135 QYSCR----LQNLDGTCADSYSFVFSRDPNGLTPETRRLVRQRQEELCLERQYRWIEHNG 190
□ 6978523    131 VYSCTTFFWFFHVD------YVWILGRNPY-LPPE                          158
□ 18543345   141 VF                                                          142
□ 6981430    131 LFS                                                         133
□ 27684523   295 PVNCR----VLFLATALALSVGFILT                                   316
```

图 4-12　一个使用格式选项 “flat query-anchored without identities.” 的 BLAST 搜索的输出结果。这里将提供一个多序列比对的 BLAST 搜索结果。其他的格式选项也是可选的，这将允许用户观察蛋白质家族中的相似区域以及趋异区域

4.3 BLAST 算法使用局部比对搜索的策略

BLAST 搜索可以找到一个数据库中与输入的查询序列相匹配的项。全局相似性算法可以得到两个序列比对的整体最优结果，这样的算法包括 GAP 程序（参见第 3 章）最适用于寻找那种包含有大段的低相似性序列的匹配。相对来说，像 BLAST 这样的局部相似性算法可以找到相对短的比对。局部比对对于数据库搜索来讲是一种很有用的方法，因为很多查询序列拥有一些结构域、活性位点或其他模体这样一些和其他蛋白具有局部而不是全局相似性的部分，而且典型的情况是数据库中拥有一些 DNA 和蛋白质序列片段，它可以局部地与一个查询序列比对上。

4.3.1 BLAST 算法组成部分：列表、扫描、延伸

BLAST 搜索算法可以找到一个查询序列与一个数据库序列之间的匹配关系，然后向两个方向进行延伸（Altschul 等，1990，1997）。搜索结果既包括数据库中高度相关的序列也包括边缘性相关的区域，并用一个打分图来描述查询序列与每个数据库匹配之间的相关程度。该算法可以描述为以下 3 个阶段（图 4-13）。

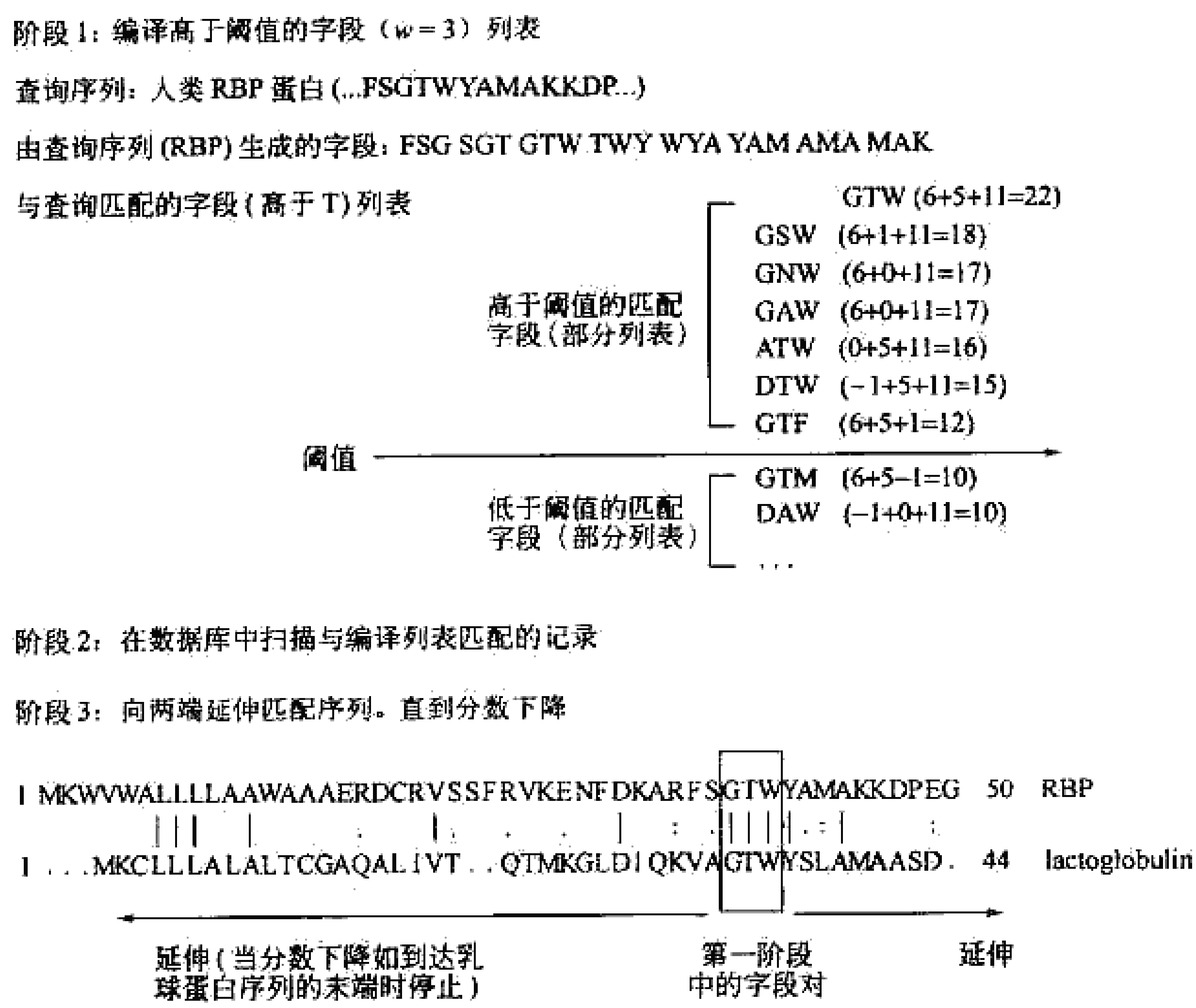


图 4-13 原始 BLAST 算法图示。在第一阶段中查询序列（如人类 RBP4）被按照指定的字段长度（如 $w=4$）分解，一个字段的列表被按照一个阈值分数（如 $T=11$）编译出来。图中列出了一些可能的字段（从 FSG 到 MAK）；在实际的 BLAST 搜索中可能有成百上千个字段被编译。对于一个指定的字段（如有 GTW 组成的查询序列的一部分）将会编译一个得分大于或者等于阈值 T（例如 11）的字段的列表。在这个例子中，显示了 9 个根据 BLOSUM62 矩阵打分的字段：其中 7 个在阈值之上，两个在阈值之下。在第二阶段，将在一个数据库中进行扫描来找到那些与编译的列表相匹配的记录。在第三阶段，数据库中的匹配将向两个方向延伸以得到一个高分片段对（HSP）。如果 HSP 的打分超过一个指定的分数阈值 S，它就将会在 BLAST 结果中输出。注意在这个特殊的例子中出发延伸步骤的字段对恰巧就是一个完全匹配（见图中方框中的残基 GTW）。然而使用阈值 T 的主要考虑就是要让那些非完全匹配的字段对也可以触发延伸步骤

① BLAST 编译一个初步的两两比对列表，成为字段对。

② 算法在一个数据库中扫描达到某个阈值分数 T 的字段对。

③ BLAST 延伸这些字段对来寻找那些超过阈值分数 S 的项，并将这些项输出给用户。这里的分数是通过打分矩阵（如 BLOSUM62）并考虑空位罚分而计算出来的。

在第一阶段中，BLAST 算法编译了一个由查询序列生成的长度固定为 w 的“字段”的列表。针对比对字段的打分将建立一个阈值 T，那些处于阈值或者阈值之上的字段将被定义为匹配；那些在阈值之下的就不再进行进一步处理。对于蛋白质搜索典型的字段大小的默认值为 3，而对于 DNA 这一默认值为 11。BLAST 用户可以更改字段大小的参数，就像下文所述的那样（参见 4.2.4.1 中⑥）。可以降低阈值 T 来得到更多的初始比对对。这样做将会增加 BLAST 查询所需要的时间同时也可能增加查询的灵敏度。

> 在 Altschul、David Lipman 及其同事们关于 BLAST 的文章中，阈值参数用 T 表示（Altschul 等，1990，1997）。在 BLAST 程序中阈值参数称为 f。

在第二阶段中，在编译了处于或高于阈值 T 的字段对列表之后，BLAST 算法将对一个数据库来进行扫描以找到匹配。这要求 BLAST 找到与编译列表中的字段相对应的数据库条目。在最初的 BLAST 运行中找到一个匹配就足够了。在现在的 BLAST 版本中，算法寻找的是两个相隔一定距离 A 的相互分离的字段（如两个没有重叠区域的字段），然后生成这两个匹配的一个无空位的延伸（Altschul 等，1997）。

> 这种双匹配的方法很大程度上提高了 BLAST 搜索的速度。与单匹配方法相比，双匹配方法平均可以生成 3 倍的匹配，但之后只需要进行原来的 1/7 的延伸操作（Altschul 等，1997）。对于参数 A，其默认值为 40。

在第三阶段中，BLAST 通过延伸这些匹配来找到称为高分片段对（HSPs）的比对。为了充分地进行高得分比对，可以进行带有空位的延伸。当打分值降低到某个阈值以下时延伸过程将结束。

> 你可以通过高级选项命令“-f 16”（或者选择其他的 f 值）来实现。或者尝试使用 NetBLAST，这是一个可以从 NCBI 网站上下载到多种操作平台上的 BLAST 版本。可以通过 FTP 在 ftp://ncbi.nlm.nih.gov/blast/network/netblast 上找到。当安装了 FTP 之后，用想要查询的序列建立一个文件（例如 rbp4.fasta）将它保存到硬盘上，然后使用命令“$ blastcl3-p blastp-d nr-i rbp4.fasta-o output1-f 11”。这个命令指定了所使用的程序（blastp）、要搜索的数据库（nr）、输入的查询文件（rbp4.fasta）、输出文件（这里为 output1）、还有最后的阈值 f。搜索请求将通过 ftp 发送到 NCBI，并将很快得到返回结果。

总起来说，BLAST 算法的主要策略是将蛋白质或 DNA 查询序列和每个数据库记录进行比较，来形成比对对（HSPs）。作为一种启发式的算法，BLAST 被设计来提供高速度和高的灵敏度。当增大阈值参数时，搜索的速度将提高，但是将得到比较少的匹配，以后数据库中那些与查询序列关系较远的匹配可能会被忽略掉。当降低阈值参数时，搜索过程将变慢很多，但会得到更多的字段匹配，从而提高了灵敏度。

> 注意使用 BLAST 进行 DNA 数据库搜索，阈值并不是用在与字段相关的地方，取而代之的是算法要求的完全的字段匹配。减小字段的长度可以有效地降低阈值分数，因此减小字段长度可以得到较慢但更为准确的搜索。

可以在一个 blastp 搜索中将 f 参数由其默认值（11）改为一个极小（5）或极大的值（16）的方法来说明不同阈值水平的影响。结果是明显的（表 4-3）。对于默认阈值 11，数据库中有 1.3 亿个匹配以及 500 万个延伸。而当阈值为 5 时，有 22 亿个数据库匹配和 5.9 亿个延伸。当阈值上升到 17 时，就只有 1200 万个匹配和 6.1 万个延伸了。在低的阈值下成功

表 4-3 在 blastp 搜索中改变阈值的效果

这3个搜索都是使用NCBI blastcl3程序(NetBLAST),以RBP4蛋白(NP_006735)作为查询项所进行的。当阈值参数(f)改为5时,得到了超过20亿的匹配和5.89亿个延伸,最终得到了比使用高的阈值17更多的允许带空位的HSP(146个)。

项　目	$f=11$(默认值)	$f=5$	$f=17$
数据库中序列的个数/个	1046476	1046476	1046476
在数据库中找到匹配的个数/个	129839417	2200945350	12002487
延伸的个数/个	5198652	589935555	61838
成功延伸的个数/个	8377	13145	1117
允许带空位的HSP的个数/个	145	146	93

延伸的数目也更多。使用默认值与使用低的阈值相比，由于带空位的 HSP 的数目差不多，所以搜索的最终结果并没有明显的差别。而对于使用高阈值的搜索有一些匹配被忽略掉了。这进一步说明了使用低阈值参数可以得到虽然速度较慢但是却更为精确的搜索。这种在灵敏度与速度之间的权衡是 BLAST 算法的核心。

4.3.2 BLAST 算法：局部比对搜索的统计学和 E 值

我们关心的是一个 BLAST 搜索的统计显著性，因为要量化地来衡量比对是否得到了显著性的匹配或者它们是否只是随机期望发生的事件。BLAST 程序提供了这样一种统计（Altschul 和 Gish，1996）。

对于全局搜索曾经介绍过一个拼凑的例子（参见第 3 章），在那里一个序列被打乱拼凑很多次并产生一个 Z 值（对于平均值的标准偏差的数值）表示比对得分。然而全局比对序列的分布是不知道的，得分的分布形状也是不清楚的。在某些情况下，分数可能不是随机分布的。因此，对于全局搜索的统计分析的了解是很少的，说一个比对得分的概率与随机期望有显著的不同也是不正确的。

对于局部比对（包括 BLAST 搜索），已经发展出了更加严格的统计检验方法（Altschul 和 Gish，1996；Altschul 等，1990，1994，1997；Pagni 和 Jongeneel，2001）。前面已经描述了如何以 HSP 来分析两个蛋白质序列间局部的、无空位的比对。使用替代矩阵可以对每一个比对残基对给出一个指定的概率值，从而对整个比对给出一个总的分数。当用查询序列与一个序列长度统一的随机序列的数据库进行比对时，将其分数进行作图后其形状是一个极值分布图（参见图 4-14，图中将它与正态分布做了比较）。正态分布（或称高斯分布）形成了我们熟悉的对称钟形曲线。极值分布则向右倾斜，并拖着一条按照 x 衰减（而不是像正态分布那样按照 x^2 衰减）的尾巴。这一分布的性质对于我们对 BLAST 的统计学的了解是至关重要的，因为这将使我们可以估计一个搜索的最高得分（例如处在分布上右手一侧尾巴上的值）随机出现的可能性。使用在线执行的 Smith-Waterman 算法、以 RBP4 蛋白为查询序列对一个蛋白质数据库进行的一个 BLAST 搜索为我们展示了极值分布的情形（图 4-15）。

由于正态分布按照 x^2 快速收尾，如果尝试用正态分布来描述一个 BLAST 搜索结果的显著性（如通过估算搜索结果处于平均值以上多少个标准差的地方）将会倾向于过高地估计比对的显著性。

下一步将考察极值分布以得到一个描述特定 BLAST 分数随机发生的可能性的公式［式（4-5）］。

图 4-14 和图 4-15 所示的极值分布的图形可以由两个参数来描述：特征值 u 和延迟常数 λ。极值分布有时也根据 1958 年第一个对它进行描述的人的名字而称为 Gumbel 分布，Altschul 和他的同事们（1994，1996）以及一些其他人（Pagni 和 Jongeneel，2001）都曾对于

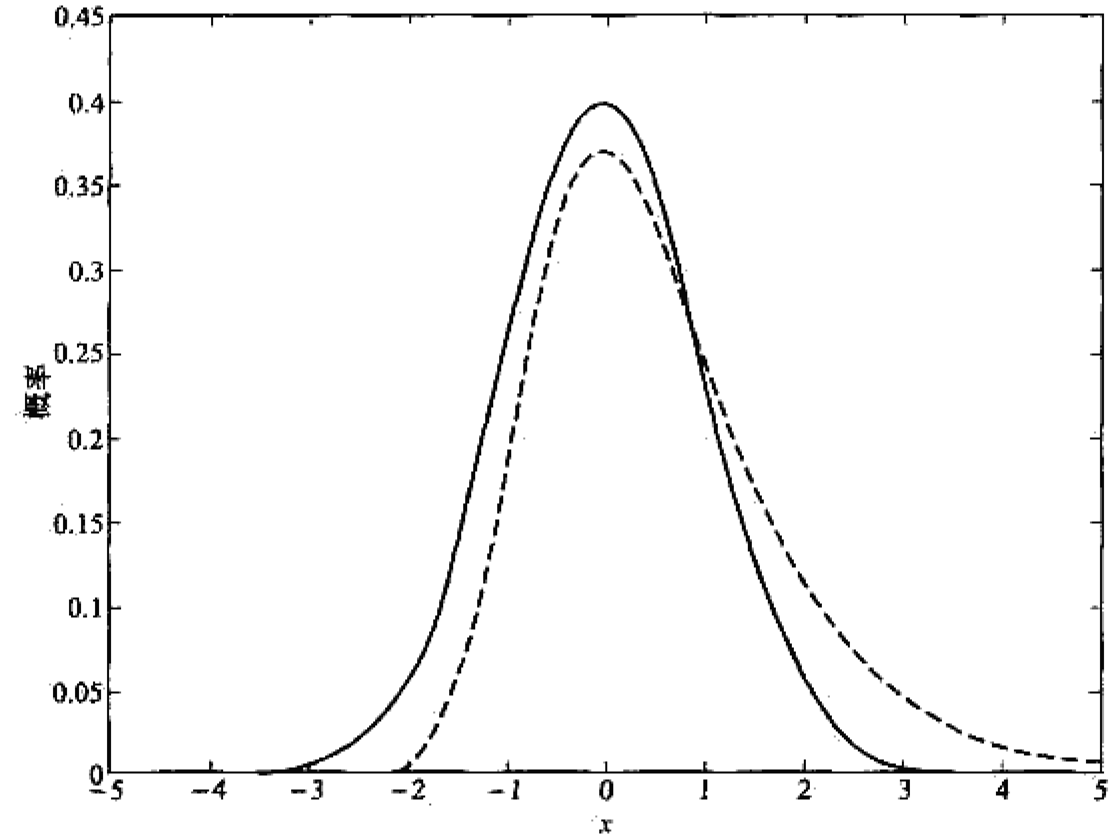


图 4-14 正态分布（实线）与极值分布（点线）的比较。将一个查询序列与一系列统一长度的随机序列的比较通常会得到符合极值分布（而不是正态分布）的分数。每条曲线下的积分面积均为1。对于正态分布，均值（μ）作为中心位于零点。获得某个分数 x 的概率 Z 由 x 到均值相距的标准偏差（σ）的个数来表示：$Z=(x-\mu)/\sigma$。与正态分布相比，极值分布是不对称的，向右倾斜。符合公式 $f(x)=(e^{-x})(e^{-e^{-x}})$。极值分布的形状由特征值 u 和延迟参数 $\lambda(u=0, \lambda=1)$ 决定

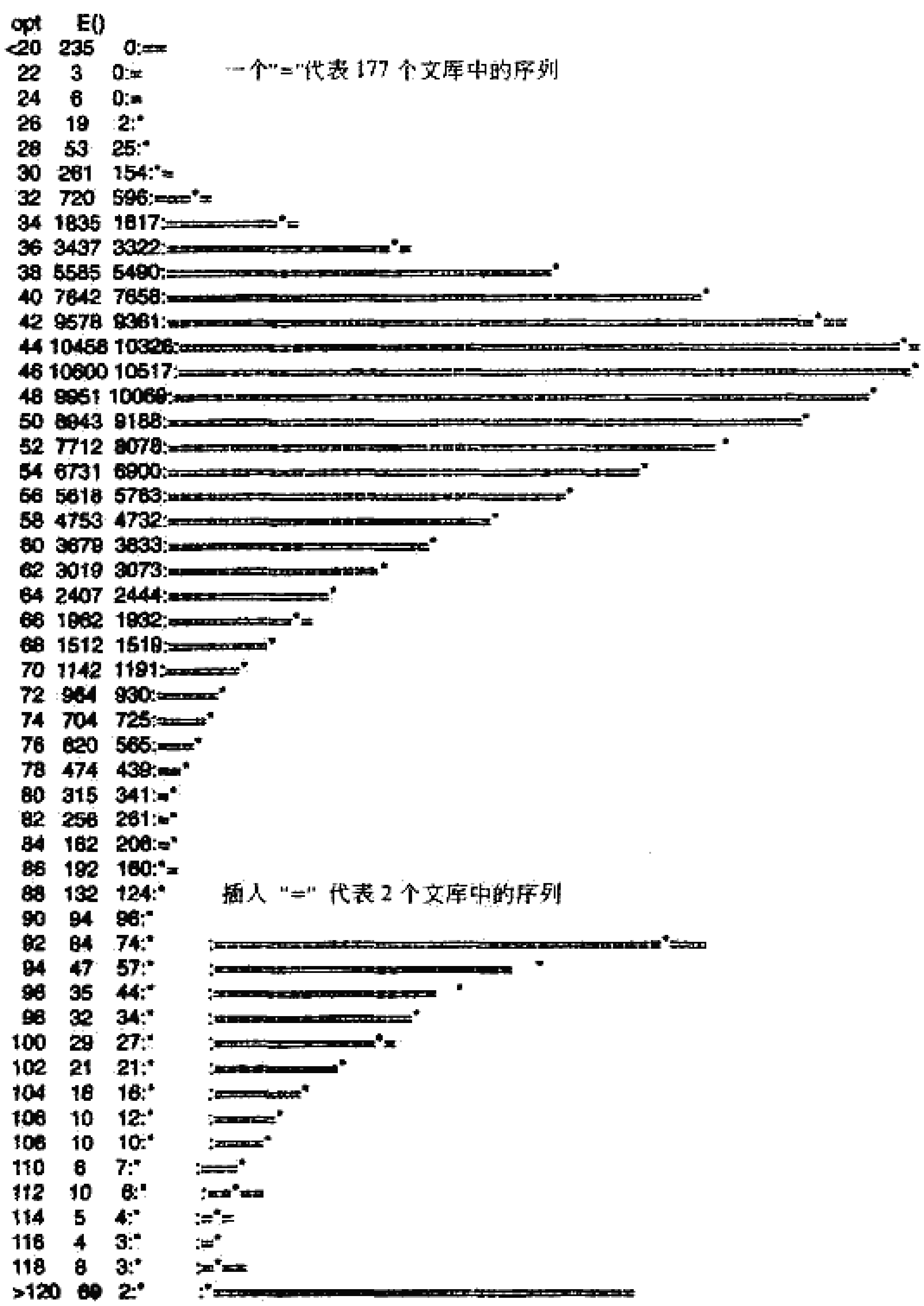


图 4-15 一个查询（RBP4）对于一个蛋白质数据库的搜索得分的分布呈现为极值分布。该搜索是使用日本网络服务器上的 DNA 数据库（http://www.ddbj.nig.ac.jp）中采用 Smith 和 Waterman 局部比对算法的 SSEARCH 程序进行的。归一化了的相似性得分（由符号"="表示）的分布与极值分布（由"*"表示）相吻合

极值分布在BLAST中的应用进行过综述。对于两个随机序列 m 和 n，得分 S 的累积分布函数可由下面公式描述：

$$P(S<x)=\exp[-e^{-\lambda(x-u)}] \tag{4-1}$$

特征值 u 与分布的最大值相关，但它不是均值 μ。

为了使用该式，需要知道（或者估算出）参数 u 和 λ 的值。对于无空位比对，参数 u 依赖于所比较的序列的长度，定义为：

$$u=\frac{\ln Kmn}{\lambda} \tag{4-2}$$

式中，m 和 n 分别表示相比较的两个序列的长度；K 为常数。联立式（4-1）和式（4-2），表示随机观察到一个等于或大于 x 的打分的概率公式为：

$$P(S\geqslant x)=1-\exp(-Kmne^{-\lambda x}) \tag{4-3}$$

我们的目标是求得一个对整个数据库的BLAST搜索得到某个结果单独随机发生的可能性。得分至少为 x 的无空位比对的数量由参数 $Kmne^{-\lambda x}$ 来描述。在数据库搜索中，m 和 n 分别指查询序列的长度（以残基数表示）和整个数据库的长度。乘积 mn 定义了搜索空间的大小。搜索空间表示了查询序列可以和数据库中任意序列进行比对的所有位置。因为一个序列的末端一般不太可能出现在一个均值化长度的比对中。BLAST算法计算有效的搜索空间时要在 m 和 n 上减去一个比对的平均长度 L（Altschul 和 Gish，1996）：

$$\text{有效搜索空间}=(m-L)(n-L) \tag{4-4}$$

在图4-16中将可以看到BLAST是如何使用这里定义的搜索空间的。

描述随机条件下得分为某个特定值 S（或更好）的HSP的期望数的公式为：

$$E=Kmne^{-\lambda S} \tag{4-5}$$

式中，E 表示期望值，即在一次数据库搜索中随机条件下期望发生的得分等于或优于 S 的不同比对的数目。这就提供了一个对于BLAST搜索中假阳性结果的估计。由式（4-5）可以看出 E 值与得分和用来度量打分系统的参数 λ 有关。同时 E 值也与查询序列的长度和数据库的长度有关。参数 K 是搜索空间的一个度量因子。参数 K 和 λ 是由 Karlin 和 Altschul 提出的（1990），因此经常被称作 Karlin-Altschul 统计量。

式（4-5）在网页 http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html#head2 中有相关的描述。

式（4-5）具有以下几个重要特点。

① 随着 S 的增加 E 值呈指数下降。分数 S 反映了每对比较的相似性并在一定程度上取决于对打分矩阵的选择。较高的 S 值对应着较好的比对；在图4-7、图4-8和图4-10中看到BLAST的结果是按照得分排序的。因此，在BLAST搜索中的一个高得分对应着一个低的 E 值。当 E 值接近零时，一个比对随机发生的可能性也就会接近零。我们将在下面将 E 值和概率值（P）联系起来。

② 比对一个随机的氨基酸对的期望分数必须是负的。否则，两个长的比对将累积很大的正分数，从而使它们即使实际上不是但也会表现为显著的相关。

③ 数据库的大小以及查询序列的长度将影响某个特定比对随机发生的可能性。考虑一个 E 值为1的BLAST结果，这个值表示对于在这样一个特定大小的数据库中只能期望随机出现一个匹配。如果数据库有原来的2倍那么大，则随机找到一个分数等于或大于 S 的可能性就会是原来的2倍。

④ 式（4-5）中包含的理论是根据无空位比对推导出来的，然而它也适用于带空位的局

部比对的情况（如一个 BLAST 搜索的结果）。对于带空位的比对，λ 和 K 不能被分析和计算出来（就如它们在无空位比对中那样），但作为替代，它们必须通过模拟来估计。

4.3.3 用比特分数来说明原始分数的意义

一个典型的 BLAST 输出中报告了 E 值和分数。共有两种分数：原始分数和比特分数。原始分数是由所选择的替代矩阵和空位罚分参数计算得到的。比特分数 S' 是由原始分数通过用描述指定打分系统的统计变量进行归一化处理后得到的。因此，从不同比对，甚至是使用不同打分矩阵的不同 BLAST 搜索的比特分数都是可以相互比较的。一个 BLAST 搜索的原始分数必须对其参数（如被查询数据库的大小）进行归一化。原始分数与比特分数的关系式为：

比特分数显示在 BLAST 输出窗口右侧的一个栏内，在每个数据库匹配的相应 E 值的旁边。

$$S' = \frac{\lambda S - \ln K}{\ln 2} \tag{4-6}$$

式中，S' 为比特分数，具有一套标准单位。E 值与一个给定的比特分数的关系为：

$$E = mn \times 2^{-S'} \tag{4-7}$$

为什么要使用比特分数？首先，原始分数是无单位的，并且自身没有什么实际意义。比特分数表明了使用的打分系统并包含了一个比对对的内在信息内容。因此它们就使得不同数据库搜索之间，即使是使用了不同的打分矩阵，也可以进行比较。第二，如果你知道了搜索空间的大小（BLAST 算法中使用的是有效搜索空间的大小，详见前文所述）、m 和 n 的乘积，通过比特分数还可以知道 E 值的大小。

4.3.4 BLAST 算法：E 值与 P 值间的关系

P 值是得分为所要求的分值比对或更好的比对随机发生的概率。它是将观测到的比对得分 S，与用同样长度和组成的随机序列作为查询序列进行的数据库搜索进行比较得到的 HSP 得分的期望分布，联系起来计算的。显著性最高的是那些接近 0 的 P 值。P 值和 E 值是反映比对显著性的两种不同方式。找到一个具有给定 E 值的 HSP 的概率为：

$$P = 1 - e^{-E} \tag{4-8}$$

表 4-4 列出了一些 E 值及与之对应的 P 值。虽然 BLAST 结果中只列出 E 值而不是 P 值。这两个值基本上是相同的，特别是在与好的数据库匹配相对应的非常小的值的情况下，在 P 值中 0.99326205 对于 0.99995460 的差别，在 E 值中就对应于更容易理解的 5 对于 10 的差别了。

表 4-4 使用式 (4-8) 计算的 E 值到 P 值的关系

小的 E 值(0.05 或更小)与 P 值更接近的对应			
E	P	E	P
10	0.99995460	0.1	0.09516258
5	0.99326205	0.05	0.04877058
2	0.86466472	0.001	0.00099950
1	0.63212056	0.0001	0.0001000

一些 BLAST 服务器（如欧洲分子生物学实验室提供的一些，参见第5章）在输出结果中使用 P 值。

传统上使用一个低于 0.05 的 P 值来定义统计的显著性（也就是拒绝一个你的查询序列与任何数据库序列都不相关的零假设）。如果零假设是真的，那么所有随机比对中有 5%将会得到明显的显著性分数。因此一个等于或小于 0.05 的 E 值可能被认为是显著的。

也可以通过保守的校正来达到 E 值的要求。当你对一个有 100 万条蛋白的数据库进行搜索时，一些得到高分的比对可能将是随机发生的。由于 E 值表明局部比对的统计显著性，一些搜索者就理所当然地考虑通过将显著性水平 α 从 0.05 调整到某个更低的值。NCBI 建议的一种尝试是定义一个归一化的分数 S'，如果它超过了 $\lg(N/E)$ 就认为是具有统计显著性的，这里 N 为搜索空间的大小，E 为期望值。

参见 http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/rules.html 中关于显著性分数的说明。

另一种尝试是通过校正 α，将其除以数据库比对的数目，如 100 万（因此 $\alpha=0.05/10^6$，或者 5×10^{-8}）。这是一种针对表明在一个查询序列对一个具有很大数量序列的数据库的搜索中找到强匹配的可能的保守校正。在分析一个完整的微生物基因组时，BLAST 或者 FASTA 搜索的 E 值如果低于 10^{-4}（Ferretti 等，2001）或者低于 10^{-5}（Chambaud 等，2001；Tettelin 等，2001；Ermolaeva 等，2001）则被认为是显著的。在人类基因组组织分析中，Smith-Waterman 比对使用的 E 值的阈值为 10^{-3}，blastn 使用的阈值为 10^{-6}（国际人类基因组测序组织，2001）。这些不同的值反映了不同水平的 P 值。我们将在第 6 章和第 7 章（基因表达和芯片数据分析）中遇到基于多重检验来校正显著性水平的方法。

4.3.5 BLAST 中的空位比对

一个空位就是在一个比对中为补偿一个序列对另一个相关序列中的插入和缺失（参见第 3 章）而引入的一个空格。既然一个单独的突变事件就可能造成多于一个残基的插入或缺失，那么一个空位的出现就比空位的长度产生更多的显著性。因此引入空位的罚分是很重的，反而空位中的每一个残基分属一个较轻的罚分。为了阻止在一个比对中积累过多的空位，引入一个空位时，就要在比对得分中扣除一定分值（空位罚分）。延伸一个空位来包括更多的核酸或者氨基酸也会在一个比对打分中进行罚分。

典型空位分数是通过空位开启罚分 G 和空位延伸罚分 L 的和计算出来的。对于一个长度为 n 的空位，空位罚分为 $G+Ln$。典型的空位罚分选择 G 为 10～15，L 为 1～2。这种引入一个空位的罚分要远远高于延伸一个空位的罚分的罚分方式被称为嫡亲空位罚分。

4.3.6 将 BLAST 搜索进行到底

每种 BLAST 算法（包括 BLAST 两个序列）都在结果输出页的底端提供一个搜索统计量的总结。图 4-16 中的一个例子显示了 blastp 搜索的参数（相同的值也显示在表 4-3 中）。在本章最后的几个问题中，我们将试着改变一个 BLAST 搜索的一些参数。这个统计量的总结提供了一种评价这些参数对于搜索过程的影响的途径。

```
           Database: All non-redundant GenBank CDS
           translations+PDB+SwissProt+PIR+PRF
             Posted date:  Jul 27, 2002 12:33 AM
1 ──→      Number of letters in database: 327,842,976
2 ──→      Number of sequences in database:  1,037,804

         Lambda      K        H
           0.321    0.136    0.431

         Gapped
         Lambda      K        H
           0.267    0.0410    0.140

3 ──→    Matrix: BLOSUM62
         Gap Penalties: Existence: 11, Extension: 1
4 ──→    Number of Hits to DB: 128,982,822
         Number of Sequences: 1037804 ←──────────────── N
         Number of extensions: 5166116
         Number of successful extensions: 8352
         Number of sequences better than 10.0: 142
         Number of HSP's better than 10.0 without gapping: 53
         Number of HSP's successfully gapped in prelim test: 89
         Number of HSP's that attempted gapping in prelim test: 8257
         Number of HSP's gapped (non-prelim): 145
         length of query: 199 ←───────────────────────── m
         length of database: 327,842,976 ←──────────── n
         effective HSP length: 115 ←──────────────────── L
         effective length of query: 84 ←────────────── m-L
         effective length of database: 208,495,516 ←── n-NL
         effective search space: 17513623344 ←─────────────── (m-L)(n-NL)
         effective search space used: 17513623344
5 ──→    T: 11
6 ──→    A: 40
         X1: 16 ( 7.4 bits)
         X2: 38 (14.6 bits)
         X3: 64 (24.7 bits)
         S1: 41 (21.9 bits)
         S2: 68 (30.8 bits)
```

图 4-16 BLAST 搜索的统计量。使用 RBP4 作为查询序列的 blastp 搜索的输出页底端部分显示了数据库的名称，数据库（箭头 1）中残基（"letter"）的数目，序列（箭头 2 和字母 N）的数目。接着，λ、K 和 H（熵）是无空位比对的值或空位比对的估计值。在矩阵（箭头 3）和空位罚分设定值的下面，记录了数据库匹配的数目（箭头 4）以及延伸、成功延伸和比期望值更好的 HSP 的数目。查询的长度（m）和数据库的长度（n）已被校正过了（［见式（4-4）］。有效的搜索空间是由查询的有效长度与数据库的有效长度相乘得到的。阈值参数（箭头 5）和触发一个延伸事件的两个独立的匹配之间间隔的长度（箭头 6）也显示了出来。此图源自 Pagni 和 Jongeneel（2001），已获使用许可

4.4 BLAST 搜索的一些策略

4.4.1 一般性概念

BLAST 搜索是一种用来对蛋白质或者 DNA 序列数据库进行搜索的工具。我们已经介绍了它的基本过程：最根本的就是要确定想要回答的问题，要输入的 DNA 或蛋白质序列，想要搜索的数据库，以及想要使用的算法。现在将要讨论一些关于 BLAST 搜索策略的基本原则（Altschul 等，1994）。我们将用脂类运载蛋白和 HIV-1 pol 的搜索来说明这些问题。关键的问题包括如何评估 BLAST 搜索结果的统计显著性和当搜索得到了太少或太多的信息时应该如何来修改 BLAST 程序的可选参数。可以用 RBP4 的 DNA（NM _ 006744）或蛋白（NP _ 006735）序列进行的所有类型的搜索总结在图 4-17 中。

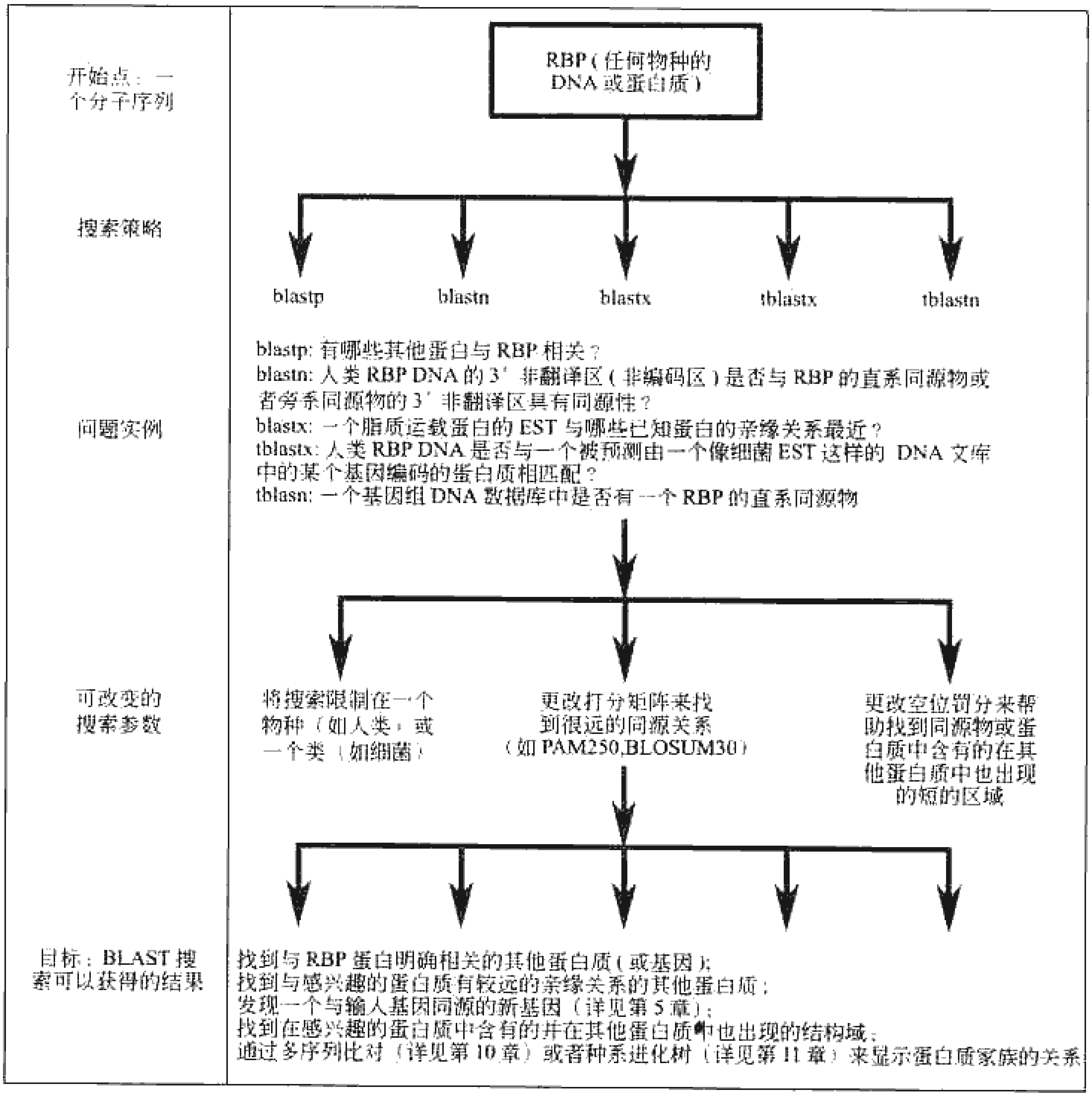


图4-17 BLAST搜索策略总图。关于BLAST搜索可以提出上百个问题，包括从刻画一个物种的基因组到评价一个单个基因的序列多态性

4.4.2 BLAST搜索的原则

4.4.2.1 如何评价你的搜索结果的显著性

几组搜索的一个一致发现是使用像期望值这样的统计分数来替代依靠一致性百分比（或相似性百分比）会降低数据库搜索的错误率。

当进行一个BLAST搜索时，哪些数据库匹配是正确的？为了回答这个问题，首先定义一个与查询序列具有同源性（由一个共同祖先演化而来）的真阳性数据库匹配。同源性是基于由搜索结果的统计评估支持的序列相似性推出的。指认基因或者蛋白质之间的相似性的问题并不能单靠序列分析解决：还需要使用生物学上的标准来支持同源性的推导。可以用对于蛋白质的结构和功能的评估来补充BLAST的结果。真正有亲缘关系的蛋白质的序列也可以有很大的差别，即使是那些有相近三维结构的蛋白质。因此我们认为数据库搜索（和蛋白质比对对）可能会得出一定数量的假阴性匹配。脂类运载蛋白家族的很多成员［如RBP4和气味分子结合蛋白（OBP)］只有非常有限的序列一致性，但它们的三维

结构关系却是非常近的，而且它们作为疏水性配体的载体的功能也被认为是相同的。

考虑一个使用人类 RBP4 蛋白作为查询项并限制在 nr 数据库中的人类蛋白的 blastp 搜索。得到 21 个结果。评估这些结果的第一个方法就是查看 *E* 值列表（图 4-18）。*E* 值得分最好（从 10^{-105} 到 10^{-13}）的 9 个条目都以 RBP 命名。这种冗余的发生是由于比对中包括了亲缘关系非常接近的几种 RBP 蛋白。在某些情况下，比对可能包含了数据库中比对序列一致而比对区域以外的序列却不同的 RBP 序列（例如蛋白质序列中的一个片段可能在数据库中堆积）。查看第 9 个条目（*E* 值为 3×10^{-13}）的比对情况，表明是一个很短的只有 36 个氨基酸的片段比对上了全长的 RBP 蛋白（图 4-19）。下一个数据库匹配是载脂蛋白 D（NP _ 001638）。RBP 片段和载脂蛋白 D 有非常相近的期望值，但在这个例子中它们与 RBP 的相似比例是不同的（94%对 31%；见图 4-19）。在任何 BLAST 搜索结果的分析中，仔细查看比对及其分数是重要的。

Sequences producing significant alignments:		Score (bits)	E Value	
gi|5803139|ref|NP_006735.1|	retinol-binding protein 4, plas...	378	e-105	L
gi|20141667|sp|P02753|RETB_HUMAN	Plasma retinol-binding pro...	375	e-104	L
gi|230284|pdb|1RBP|	Retinol Binding Protein >gi|493897|pdb...	371	e-103	□
gi|4558179|pdb|1QAB|E	Chain E, The Structure of Human Retin...	363	e-100	□
gi|7770173|gb|AAF69622.1|AF119917_30	PRO2222 [Homo sapiens]	324	6e-89	L
gi|296672|emb|CAA26553.1|	RBP [Homo sapiens]	207	9e-54	L
gi|5419892|emb|CAB46489.1|	RBP (aa 101-172) [Homo sapiens]	149	3e-36	L
gi|2895204|gb|AAC02945.1|	mutant retinol binding protein [H...	90	2e-18	L
gi|2895206|gb|AAC02946.1|	mutant retinol binding protein [H...	73	3e-13	L
gi|4502163|ref|NP_001638.1|	apolipoprotein D precursor [Hom...	55	5e-08	L
gi|619383|gb|AAB32200.1|	apolipoprotein D, apoD [human, pla...	55	5e-08	
gi|1246096|gb|AAB35919.1|	apolipoprotein D; apoD [Homo sapi...	43	3e-04	L
gi|223373|prf||0801163A	complex-forming glycoprotein HC	37	0.012	
gi|21730336|pdb|1IW2|A	Chain A, X-Ray Structure Of Human Co...	36	0.041	□
gi|4884164|emb|CAB43305.1|	hypothetical protein [Homo sapiens]	35	0.048	L
gi|130701|sp|P09466|PAEP_HUMAN	Glycodelin precursor (GD) (P...	35	0.048	L
gi|4502067|ref|NP_001624.1|	alpha-1-microglobulin/bikunin p...	35	0.077	L
gi|4557393|ref|NP_000597.1|	complement component 8, gamma p...	34	0.16	L
gi|4505583|ref|NP_002562.1|	progestagen-associated endometr...	32	0.54	L
gi|17468008|ref|XP_070794.1|	similar to Glycodelin precurso...	29	3.1	L

图 4-18 以人类 RBP 蛋白为查询序列的 blastp nr 搜索的结果（输出结果限制在人类蛋白质中）。这里有 21 个匹配，通过对 *E* 值的观察表明除了 RBP 本身，该搜索还找到了一些可能的真正同源物。补体成分 8 和一个孕激素相关蛋白是否可能与 RBP 同源

列表的下面可看到一个孕激素相关子宫内膜蛋白（也称为胎盘蛋白 14 和孕激素相关子宫内膜 α-2-球蛋白；NP _ 002562）的数据库匹配。该比对具有一个高的、非显著性的 *E* 值（0.54）和一个低的比特分数，该蛋白仅仅与 RBP 蛋白在一段 107 个氨基酸残基的区域上具有 24%的氨基酸一致性，比对中还包括四个空位（图 4-20）。根据结果可能得出这两个蛋白不具有同源性的结论。但在这个例子中它们却是同源的。为决定两个蛋白质（或者 DNA 序列）是否同源，可以问以下几个问题。

> 我们将在第 8 章介绍模体和名签，在第 11 章介绍进化树。

① 期望值究竟是不是显著？在我们这个特定的例子里它不是。

② 两个蛋白质是不是具有近似的大小？实际上同源蛋白质并不要求具有相同的大小，两个蛋白质可能只共享有限的一个相同的结构域。然而，产生一个对两个蛋白质同源的可能性的生物学上的直觉是非常重要的。一个 1000 个氨基酸具有跨膜结构域的蛋白质不太可能

```
□>gi|2895206|gb|AAC02946.1|  mutant retinol binding protein [Homo sapiens]
          Length = 36

 Score = 72.8 bits (177), Expect = 3e-13
 Identities = 34/36 (94%), Positives = 35/36 (97%)

Query: 82   NWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGN 117
            NWDVCADMV TFTDTEDPAKFKMKYWGVASFLQKG+
Sbjct: 1    NWDVCADMVDTFTDTEDPAKFKMKYWGVASFLQKGS 36

□>gi|4502163|ref|NP_001638.1|  apolipoprotein D precursor [Homo sapiens]
 gi|114034|sp|P05090|APOD_HUMAN  Apolipoprotein D precursor (Apo-D) (ApoD)
 gi|72088|pir||LPHUD  apolipoprotein D precursor [validated] - human
 gi|178841|gb|AAB59517.1|  apolipoprotein D precursor
 gi|178847|gb|AAA51764.1|  apolipoprotein D precursor
 gi|13938509|gb|AAH07402.1|AAH07402  apolipoprotein D [Homo sapiens]
          Length = 189

 Score = 55.5 bits (132), Expect = 5e-08
 Identities = 47/151 (31%), Positives = 78/151 (51%), Gaps = 39/151 (25%)

Query: 27   VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC 86
            V+ENFD   ++ G WY + +K P         I A +S+ E           G++++LN  ++
Sbjct: 33   VQENFDVNKYLGRWYEI-EKIPTTFENGRCIQANYSLMEN---------GKIKVLNQ-ELR 82

Query: 87   ADMVGTFTDTE---------DPAKFKMKY-WGVASFLQKGNDDHWIVDTDYDTYAVQYSC 136
            AD  GT    E          +PAK ++K+ W + S          +WI+ TDY+ YA+ YSC
Sbjct: 83   AD--GTVNQIEGEATPVNLTEPAKLEVKFSWFMPS-------APYWILATDYENYALVYSC 134

Query: 137  ----RLLNLDGTCADSYSFVFSRDPNGLPPE 163
                +L ++D      ++++ +R+PN LPPE
Sbjct: 135  TCIIQLFHVD------FAWILARNPN-LPPE 158
```

图 4-19 用人类RBP4蛋白进行搜索时返回结果中的两对比对（参见图4-18的下半部）。一个只有36个氨基酸长的RBP4框架就可以得到与RBP4和载脂蛋白D的较长的匹配相同的得分和期望值。这个结果突出了考察BLAST搜索中每个比对的必要性

```
□>gi|4505583|ref|NP_002562.1|  progestagen-associated endometrial protein (placental protein 14,
          pregnancy-associated endometrial alpha-2-globulin, alpha
          uterine protein); Progestagen-associated endometrial
          protein (placental protein 14) [Homo sapiens]
 gi|190215|gb|AAA60147.1|  placental protein 14
          Length = 162

 Score = 32.0 bits (71), Expect = 0.54
 Identities = 26/107 (24%), Positives = 48/107 (44%), Gaps = 11/107 (10%)

Query: 26   RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWD- 84
            + K++ +   + +GTW++MA       + L    + A   V  T  +            +L+ W+
Sbjct: 5    QTKQDLELPKLAGTWHSMAMAT-NNISLMATLKAPLRVHITSLLPTPEDNLEIVLHRWEN 63

Query: 85   -VCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTY 130
              C +         T +P KFK+ Y  VA       ++   ++DTDYD +
Sbjct: 64   NSCVEKKVLGEKTGNPKKFKINY-TVA-------NEATLLDTDYDNF 102
```

图 4-20 人类RBP4蛋白（查询序列）与孕激素相关子宫内膜蛋白的比对。比特分数相对较低，期望值（0.49）并不具有显著性，而且在局部比对中两个蛋白质只在107个氨基酸的区段上共享24%的序列一致性。虽然如此，这两个蛋白却是同源的。它们的同源性可以被确认，因为：①两个蛋白质的大小近似相同；②它们共有一个包含一个GXW模体的载脂蛋白名签；③它们是多序列比对的（详见第5章和第10章）；④它们都是可溶、亲水和含量丰富的蛋白质，可能共同具有与载体蛋白相似的功能；⑤它们很可能共同具有一个相似的三维结构（参见第9章）

与 RBP 蛋白同源，因为绝大多数的脂类运载蛋白长度都近似 200 个氨基酸（20～25ku）。

③ 这两个蛋白质是否有共同的模体或信号序列？在这个例子中，孕激素相关子宫内膜蛋白和 RBP4 蛋白都含有一个 GXW 载脂蛋白名签。就像已经简略看到的（图 3-3 和图 3-9），它们都是一个载脂蛋白多序列比对的一个部分（并且因此它们都是同一个系统发生进化树的一部分）。

④ 这两个蛋白质是不是一个合理的多序列比对的一部分？我们将在第 5 章和第 10 章中看到这样的例子。

⑤ 这两个蛋白质是否共有一个相似的生物学功能？就像所有的载脂蛋白一样，这两个蛋白都是小的、亲水性的、含量丰富的分泌分子。

⑥ 这两个蛋白质是否具有相似的三维结构？虽然载脂蛋白序列多种多样，它们却共享一个显著的非常保守的结构。这个结构（一个杯状的盂）可以使它们将疏水性配体运输通过一个水环境区间（参见第 9 章）。

⑦ 如果一个 BLAST 搜索得到一个对另一个蛋白质的边缘匹配，以这个具有较远亲缘关系的蛋白质作为查询项再进行一次新的 BLAST 搜索。使用孕激素相关子宫内膜蛋白作为查询项进行 blastp nr 搜索可以得到一些也在 RBP4 蛋白的搜索中也检测到的蛋白质（复合体形成糖蛋白 HC 和 α-1-微球蛋白/人类 α-内切抑制双头蛋白酶）（图 4-21）。这个发现增加

```
                                                                Score    E
Sequences producing significant alignments:                     (bits) Value

gi|4505583|ref|NP_002562.1|  progestagen-associated endometr...   321   3e-88
gi|4884164|emb|CAB43305.1|  hypothetical protein [Homo sapiens]   318   1e-87
gi|130701|sp|P09466|PAEP_HUMAN  Glycodelin precursor (GD) (P...   318   1e-87
gi|182093|gb|AAA35802.1|  pregnancy-associated endometrial a...   264   3e-71
gi|17468008|ref|XP_070794.1|  similar to Glycodelin precurso...   122   2e-28
gi|482787|pir||A61166  endometrial progesterone-dependent pr...    47   1e-05
gi|22045262|ref|XP_172757.1|  similar to beta-lactoglobulin ...    46   2e-05
gi|20177973|sp|Q8WX39|MUPL_HUMAN  Putative MUP-like lipocali...    37   0.008
gi|223373|prf||0801163A  complex-forming glycoprotein HC           37   0.017
gi|4502067|ref|NP_001624.1|  alpha-1-microglobulin/bikunin p...    35   0.053
gi|6900069|emb|CAB71319.1|  putative odorant binding protein...    34   0.089
gi|7657405|ref|NP_055397.1|  odorant-binding protein 2A [Hom...    33   0.13
gi|27486042|ref|XP_172158.2|  similar to C-1-tetrahydrofolat...    33   0.16
gi|6900073|emb|CAB71321.1|  putative odorant binding protein...    32   0.27
gi|4558179|pdb|1QAB|E  Chain E, The Structure Of Human Retin...    32   0.29
gi|230284|pdb|1RBP|  Retinol Binding Protein >gi|493897|pdb...     32   0.29
gi|20141667|sp|P02753|RETB_HUMAN  Plasma retinol-binding pro...    32   0.29
gi|6900083|emb|CAB71326.1|  odorant binding protein a [Homo ...    32   0.30
gi|404390|gb|AAB27607.1|  beta-trace protein, prostaglandin ...    32   0.32
gi|5803139|ref|NP_006735.1|  retinol-binding protein 4, plas...    32   0.37
gi|190444|gb|AAA36494.1|  prostaglandin D synthase precursor...    32   0.38
gi|27498839|ref|XP_208403.1|  similar to Von Ebners gland pr...    31   0.82
gi|7657407|ref|NP_055396.1|  odorant-binding protein 2B [Hom...    30   1.6
gi|6900079|emb|CAB71324.1|  putative odorant-binding protein...    29   2.1
gi|4504963|ref|NP_002288.1|  lipocalin 1 (protein migrating ...    29   2.6
gi|296672|emb|CAA26553.1|  RBP [Homo sapiens]                      28   6.0
gi|2495324|sp|Q92826|HXBD_HUMAN  Homeobox protein Hox-B13 >g...    28   7.6
gi|5453686|ref|NP_006352.1|  homeo box B13; homeobox protein...    28   7.8
gi|18203852|gb|AAH21172.2|AAH21172  Unknown (protein for MGC...    27   9.9
gi|18314571|gb|AAH22017.1|AAH22017  Unknown (protein for MGC...    27  10.0
```

图 4-21　使用孕激素相关子宫内膜蛋白（NP_002562）作为查询项进行 blastp nr 搜索（限定在人类蛋白质的范围内）的结果。该蛋白与一些也出现在一个 RBP4 蛋白的 blastp 搜索的输出结果中的蛋白质（复合体形成糖蛋白 HC 和 α-1-微球蛋白/人类 α-内切抑制双头蛋白酶）相关。这种重叠支持了孕激素相关子宫内膜蛋白与 RBP4 蛋白确实相关的假说。使用两个可能相关的序列分别作为查询序列进行相互 BLAST 搜索经常是重要的

了我们认为RBP4蛋白和孕激素相关子宫内膜蛋白实际上是一个蛋白超家族中的同源成员的信心。如果使用孕激素相关子宫内膜蛋白进行的blastp搜索表明该蛋白质是另一个特征蛋白质家族的成员，这将很大程度上降低我们认为它是真正与RBP4蛋白相关的信心。

登录LocusLink，输入"rhodopsin"，将物种限制为人类。有超过700个的搜索结果，大部分由被认为含有7个跨膜区段的受体家族的成员组成，在第8章中将看到怎样搜索蛋白质家族。

历史上，早期的数据库搜索将得出完全出乎意料的结果。1984年，β-肾上腺素能受体被发现与视紫红质蛋白同源（Dixon等，1986）。这是令人惊奇的，因为这两种受体在功能和细胞定位上都具有明显的差异：视紫红质蛋白是一个视网膜特异性的光受体，而肾上腺素能受体被认为与肾上腺素和去甲肾上腺素结合，刺激一个级联信号传导从而引起环腺苷一磷酸的产生。蛋白质序列的比对揭示出它们共享相似的结构特点（7个预测的跨膜结构域）。现在我们高兴地看到视紫红质蛋白和β-肾上腺素能受体都是一个与配体结合启动第二信使级联反应的蛋白质超家族的原始成员。另一个令人吃惊的发现是一些转导哺乳动物细胞的病毒基因实际上是由其宿主物种中演化而来的。对人类上皮生长因子受体测序，发现它与鸟类逆转录致癌基因 *v-erbB* 同源（Downward等，1984）。还有很多揭示出意料之外的关系的数据库搜索的例子。下面是一个用HIV序列进行BLAST搜索的例子。

4.4.2.2 如何处理过多的结果

在BLAST检索中经常遇到的一种情况是太多的结果被反馈回来。许多策略可用来减少结果的数量，但要从中做出恰当的选择，必须集中注意力在要解决的问题上。

① 在"limit by Entrez query"窗口中输入"refseq"，这样所有的返回结果中都会带有一个RefSeq号。这经常可以去掉冗余的数据库匹配结果。

② 可能的话，利用生物体的种类对数据库反馈结果做出限制。这将会减少无关的信息。如果利用BLAST服务器的选项特征，系统将执行同样规模的搜索。相反地，如果你选择了一个针对特定生物体的数据库，搜索的速度经常会大大提高。第5章将介绍一些针对特定生物体的BLAST服务器。

③ 适当的话，利用序列的一部分进行搜索。利用独立的结构域序列就可以进行多结构域蛋白的检索。如果你在研究HIV-1 pol，可能会对整个蛋白感兴趣或是对一个特殊的部分感兴趣，例如反转录酶结构域。

④ 调整打分矩阵使其更恰当地体现你的query和数据库匹配项之间的相似度。

⑤ 调整期望值：降低 E 值可减少反馈回来的数据库中的匹配项。

4.4.2.3 如何处理过少的结果

很多基因或是蛋白在数据库中没有或是只有极少的匹配项。当新的微生物基因组测序完成时，预测得到的蛋白中有一半不和其他任何蛋白相匹配（见第14章）。显而易见有这样一些用于提高BLAST搜索得到的数据库匹配项数目的策略：去掉Entrez限制，提高期望值，尝试更高PAM值或更低BLOSUM值的打分矩阵。还可以从很多种不同的附加数据库中搜索。在NCBI站点上，可以搜索所有可提供的数据库（例如HTGS和GSS）。很多基因组序列中心保存着不同的数据库可通过BLAST进行搜索。以上内容将在第5章（高级BLAST检索）中详细描述。另外，有很多数据库检索工具比BLAST更为灵敏。其中包括position-specific scoring matrix（PPSMs）和hidden Markov models（HMMs），这些在第5章和第10章中都有描述。

4.5 多结构域蛋白（HIV-1 pol）的 BLAST 检索

HIV-1 的 pol 蛋白（NP _ 057849）有 1003 个氨基酸残基并包含有多个不同的蛋白酶、反转录酶和整合酶（integrase）结构域。因此它是一个多结构域蛋白的例子。图 4-22 总结了对于这样的一个病毒蛋白可进行的搜索的种类。

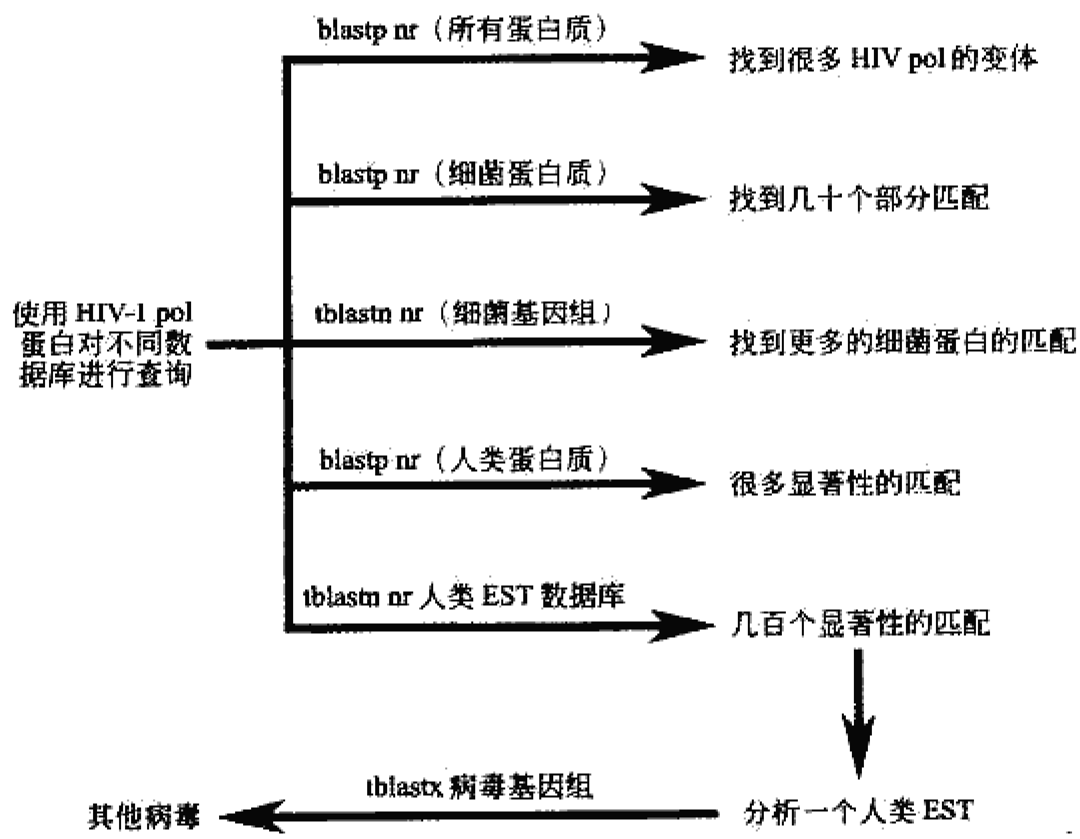


图 4-22 以 HIV-1 pol 蛋白开始的 BLAST 搜索总图。通常为研究一个特定基因、蛋白或物种可以进行一系列的 BLAST 搜索。BLAST 搜索返回的数据库匹配结果的数量可以从一个到上千个，这完全取决于查询序列、数据库和搜索参数本身的特点

当用这个蛋白对去冗余蛋白数据库进行 BLAST 检索时，先输入 RefSeq 序列号（NP _ 057849），再点击提交。程序报告已检测到可能的保守结构域，示意图指出各个结构域的方位（图 4-23）。点击这些结构域中的任意一个，将链接到 NCBI 保守结构域数据库或链接到 Pfam 和 SMART 数据库（参见第 5 章和第 8 章）。继续 BLAST 检索，将会看到有 500 个检索结果，全都有极低的期望值，并且都符合 HIV 和猿免疫缺陷病毒（SIV）的 pol 蛋白匹配项（图 4-24）。将输出格式变为“query-anchored without identities”可察看到这些 pol 蛋白的保守性（图 4-25）。

要想对 pol 蛋白在整个生命树中的分布了解得更多，我们可能会问有哪些细菌蛋白和 HIV-1 pol 蛋白相关。再次利用 NP _ 057849 作为 query 进行 blastp 检索，这次从生物体的下拉菜单中选择“细菌”。这里，BLAST 检索结果的图表总结是大有帮助的，它能使你看到，与已知细菌序列匹配的大多数都是病毒 pol 蛋白的两个结构域，它们分别对应于 pol 蛋白第 600 到第 700 以及第 800 到第 900 的氨基酸残基（图 4-26）。将输出结果与 HIV-1 pol 的结构域构成进行对比后发现这两个与细菌蛋白相匹配的病毒蛋白结构域是 RNase H 和整合酶。的确，与病毒 pol 匹配的细菌蛋白包括核糖核酸酶 H 和转座酶 [图 4-27(a)]。比对结果显示这些病毒蛋白和细菌蛋白是同源的，有跨越 166 个氨基酸残基的 30% 的相似性。

现在来关注可能与 HIV-1 pol 同源的人类蛋白。这个检索与上面的大致相同，只需要把生物体选为人类。有趣的是在最终得到的多于 50 个的匹配项中，有许多跨越病毒 pol 的全长（图 4-28）。这些人类蛋白被注释为聚合酶 [图 4-28(b)]。这些人类基因如果

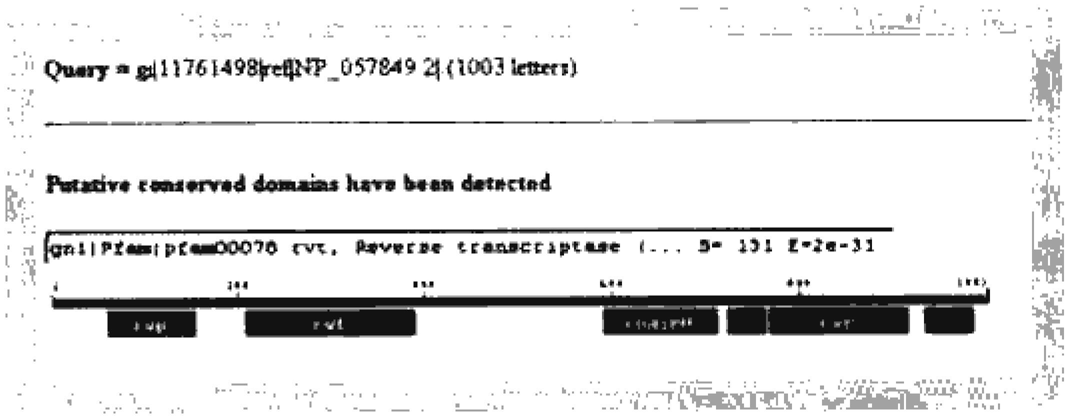


图 4-23 一个使用病毒 pol（NP_057849）进行的 blastp 搜索显示了蛋白中的保守区域。这些模块都是可以点击的，可以链接到另外的数据库（Pfam，SMART：参见第 8 章）。其中，rvp 表示反转录病毒天冬氨酸蛋白酶；rvt 表示反转录酶（RNA 依赖性 DNA 合成酶）；rnaseH 表示核酸酶 H；小的未标记框表示整合酶 DNA 结合区

```
                                                              Score    E
Sequences producing significant alignments:                   (bits) Value

gi|28872819|ref|NP_057849.4|  Gag-Pol: Gag-Pol polyprotein [...   2836   0.0
gi|11066865|gb|AAG28737.1|AF287353_1  gag-pol fusion protein...   2756   0.0
gi|3002831|gb|AAD03191.1|  gag-pol fusion polyprotein [Human...   2753   0.0
gi|19072104|dbj|BAB85751.1|  Gag-pol fusion polyprotein [Hum...   2744   0.0
gi|3002841|gb|AAD03200.1|  gag-pol fusion polyprotein [Human...   2731   0.0
gi|11095912|gb|AAG30116.1|AF286365_2  gag-pol fusion polypro...   2723   0.0
gi|3002860|gb|AAD03217.1|  gag-pol fusion polyprotein [Human...   2716   0.0
gi|3193273|gb|AAD03326.1|  gag-pol fusion polyprotein [Human...   2691   0.0
gi|3002887|gb|AAD03241.1|  gag-pol fusion polyprotein [Human...   2688   0.0
gi|3002869|gb|AAD03225.1|  gag-pol fusion polyprotein [Human...   2685   0.0
gi|3002851|gb|AAD03209.1|  gag-pol fusion polyprotein [Human...   2674   0.0
gi|3002878|gb|AAD03233.1|  gag-pol fusion polyprotein [Human...   2667   0.0
gi|25166689|gb|AAN73492.1|AF484483_1  gag-pol fusion polypro...   2647   0.0
gi|25166961|gb|AAN73736.1|AF484511_1  gag-pol fusion polypro...   2640   0.0
gi|25166629|gb|AAN73438.1|AF484477_1  gag-pol fusion polypro...   2633   0.0
gi|25166991|gb|AAN73763.1|AF484514_1  gag-pol fusion polypro...   2632   0.0
gi|25166901|gb|AAN73682.1|AF484505_1  gag-pol fusion polypro...   2630   0.0
gi|25166951|gb|AAN73727.1|AF484510_1  gag-pol fusion polypro...   2628   0.0
gi|25166731|gb|AAN73529.1|AF484488_1  gag-pol fusion polypro...   2627   0.0
gi|25167011|gb|AAN73781.1|AF484516_1  gag-pol fusion polypro...   2627   0.0
gi|25166791|gb|AAN73583.1|AF484494_1  gag-pol fusion polypro...   2627   0.0
gi|25166821|gb|AAN73610.1|AF484497_1  gag-pol fusion polypro...   2626   0.0
gi|25166711|gb|AAN73511.1|AF484486_1  gag-pol fusion polypro...   2625   0.0
gi|25166751|gb|AAN73547.1|AF484490_1  gag-pol fusion polypro...   2624   0.0
gi|25166669|gb|AAN73474.1|AF484481_1  gag-pol fusion polypro...   2622   0.0
gi|25166981|gb|AAN73754.1|AF484513_1  gag-pol fusion polypro...   2622   0.0
gi|25166801|gb|AAN73592.1|AF484495_1  gag-pol fusion polypro...   2622   0.0
gi|25166891|gb|AAN73673.1|AF484504_1  gag-pol fusion polypro...   2622   0.0
gi|25167041|gb|AAN73808.1|AF484519_1  gag-pol fusion polypro...   2621   0.0
gi|25166911|gb|AAN73691.1|AF484506_1  gag-pol fusion polypro...   2620   0.0
gi|25166841|gb|AAN73628.1|AF484499_1  gag-pol fusion polypro...   2620   0.0
gi|25166741|gb|AAN73538.1|AF484489_1  gag-pol fusion polypro...   2619   0.0
gi|25166831|gb|AAN73619.1|AF484498_1  gag-pol fusion polypro...   2619   0.0
gi|18073999|emb|CAC86564.1|  gag-pol fusion polyprotein prec...   2618   0.0
gi|25166871|gb|AAN73655.1|AF484502_1  gag-pol fusion polypro...   2617   0.0
gi|2570314|gb|AAC97575.1|  gag-Pol fusion [Human immunodefic...   2615   0.0
gi|25166679|gb|AAN73483.1|AF484482_1  gag-pol fusion polypro...   2614   0.0
gi|25166771|gb|AAN73565.1|AF484492_1  gag-pol fusion polypro...   2613   0.0
gi|25166701|gb|AAN73502.1|AF484485_1  gag-pol fusion polypro...   2613   0.0
gi|25166851|gb|AAN73637.1|AF484500_1  gag-pol fusion polypro...   2612   0.0
gi|25167071|gb|AAN73835.1|AF484522_1  gag-pol fusion polypro...   2608   0.0
gi|4324862|gb|AAD17126.1|  gag-pol polyprotein [Human immuno...   2604   0.0
gi|25167021|gb|AAN73790.1|AF484517_1  gag-pol fusion polypro...   2603   0.0
gi|25166721|gb|AAN73520.1|AF484487_1  gag-pol fusion polypro...   2603   0.0
gi|25166659|gb|AAN73465.1|AF484480_1  gag-pol fusion polypro...   2600   0.0
gi|25167001|gb|AAN73772.1|AF484515_1  gag-pol fusion polypro...   2597   0.0
```

图 4-24 使用 HIV-1 pol 进行 blastp nr 搜索得到了非常多看上去像是 HIV-1 的变体的数据库匹配，这里所有显示的 E 值都是 0，这个结果模糊了那些不是来自 HIV-1 的可能的匹配

```
□ 1_19000   294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 353
□ 28872819  294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 353
□ 11066865  294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 353
□ 3002831   294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPGATLEEMMTACQGV 353
□ 19072104  294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 353
□ 3002841   295  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 354
□ 11095912  294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPGATLEEMMTACQGV 353
□ 3002860   294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNSNPDCKTILKALGPAATLEEMMTACQGV 353
□ 3193273   294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNSNPDCKTILKALGPGATLEEMMTACQGV 353
□ 3002887   299  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 358
□ 3002869   294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 353
□ 3002851   299  RDYVDRFYKTLRAEQATQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 358
□ 3002878   294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCRTILKALGPAATLKEMMTACQGV 353
□ 25166689  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPGATLEEMMTACQGV 351
□ 25166961  292  RDYVDRFYKTLRAEQATQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 25166629  292  RDYVDRFYKTLRAEQASQEVKGWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 25166991  292  RDYVDRFYKTLRAEQASQEVKSWMTETLLVQNANPDCKTILKSLGTGATLEEMMTACQGV 351
□ 25166901  292  RDYVDRFYKTLRAEQATQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 25166951  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 25166731  291  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 350
□ 25167011  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 25166791  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLIQNANPDCKTILKALGPGATLEEMMTACQGV 351
□ 25166821  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCRTILKALGPAATLEEMMTACQGV 351
□ 25166711  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 25166751  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 25166669  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 25166981  293  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPGASLEEMMTACQGV 352
□ 25166801  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 25166891  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPGATLEEMMTACQGV 351
□ 25167041  292  RDYVDRFYKTLRAEQATQEVKNWMTETLLIQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 25166911  292  RDYVDRFYKTLRAEQASQDVKTWMTETLLVQNANPDCKTILKALGPQATLEEMMTACQGV 351
□ 25166841  290  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 349
□ 25166741  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 25166831  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 18071999  294  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPGATLEEMMTACQGV 353
□ 25166871  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCRTILKALGPGATLEEMMTACQGV 351
□ 2570314   294  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 353
□ 25166679  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 25166771  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 25166701  292  RDYVDRFYKTLRAEQASQEVKGWMTETLLVQNANPDCKTILKALGPGATLEEMMTACQGV 351
□ 25166851  292  RDYVDRFFKILRAEQATQEVKNWMTDTLLVQNANPDCKTILKALGPAASLEEMMTACQGV 351
□ 25167071  292  RDYVDRFFKTLRAEQATQEVKGWMTDTLLVQNANPDCKTILKALGPGASLEEMMTACQGV 351
□ 4324862   291  RDYVDRFFKTLRAEQATQDVKNWMTDTLLVQNANPDCKTILRALGPGATLEEMMTACQGV 350
□ 25167021  292  RDYVDRFYKTLRAEQASQEVKNWMTETLLIQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 25166721  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
□ 25166659  292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAASLEEMMTACQGV 351
```

图 4-25 使用 HIV-1 pol 蛋白（NP _ 057849）进行的一个 blastp 搜索的部分结果。这种平铺并于查询序列一起列出的结果输出格式表示出了那些被置换的氨基酸残基以及那些保留的不变氨基酸残基

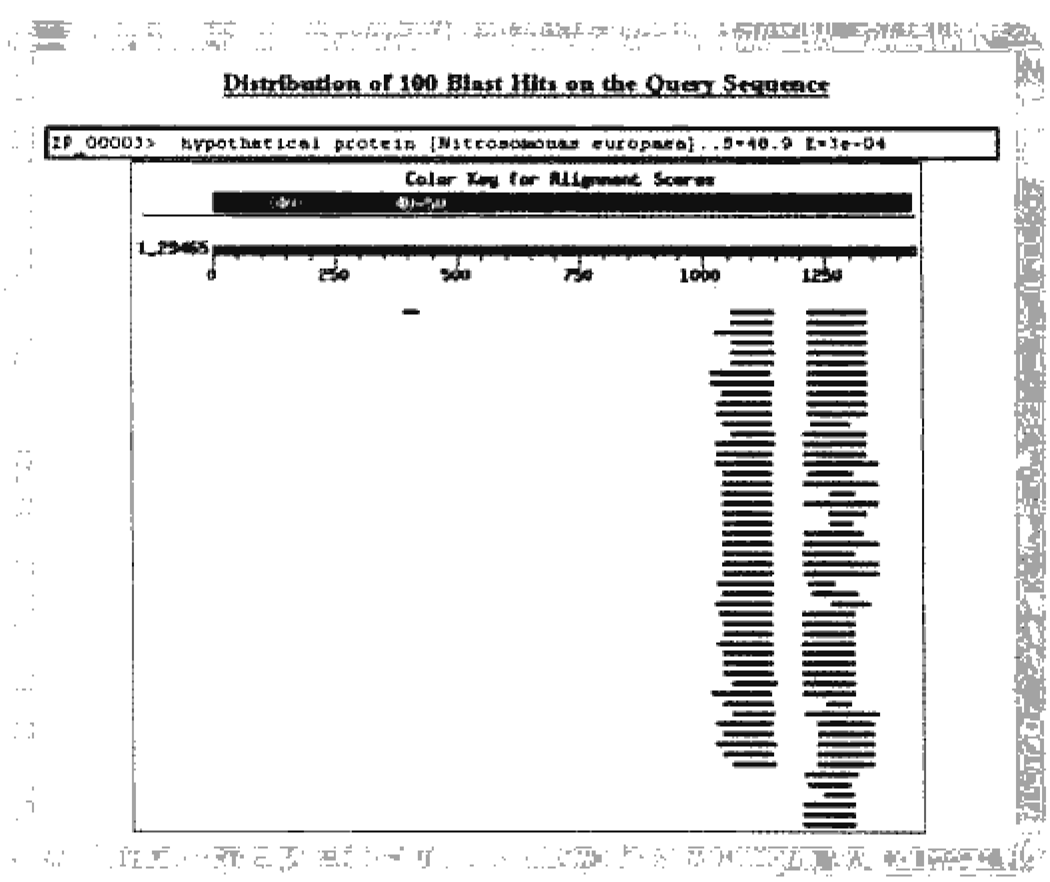


图 4-26 将搜索范围限制在细菌中时使用 HIV-1 pol 蛋白进行 blastp 搜索的结果。图形化的 BLAST 搜索结果使很容易确定 HIV-1 中那些具有细菌中同源物的区域（参见图 4-23）

(a)

```
                                                                     Score    E
Sequences producing significant alignments:                          (bits) Value

gi|22955850|gb|ZP_00003651.1|  hypothetical protein [Nitroso...        49   3e-04
gi|22954454|gb|ZP_00002255.1|  hypothetical protein [Nitroso...        49   3e-04
gi|22954317|gb|ZP_00002118.1|  hypothetical protein [Nitroso...        49   3e-04
gi|22954687|gb|ZP_00002488.1|  hypothetical protein [Nitroso...        49   3e-04
gi|22956290|gb|ZP_00004091.1|  hypothetical protein [Nitroso...        49   3e-04
gi|22955780|gb|ZP_00003581.1|  hypothetical protein [Nitroso...        49   3e-04
gi|22955002|gb|ZP_00002803.1|  hypothetical protein [Nitroso...        49   3e-04
gi|22954366|gb|ZP_00002167.1|  hypothetical protein [Nitroso...        49   3e-04
gi|22955688|gb|ZP_00003489.1|  hypothetical protein [Nitroso...        49   3e-04
gi|22956184|gb|ZP_00003985.1|  hypothetical protein [Nitroso...        49   3e-04
gi|22954280|gb|ZP_00002081.1|  hypothetical protein [Nitroso...        49   3e-04
gi|22954184|gb|ZP_00001985.1|  hypothetical protein [Nitroso...        43   0.012
gi|23103018|ref|ZP_00089511.1|  hypothetical protein [Azotob...        42   0.042
gi|23469154|ref|ZP_00124469.1|  hypothetical protein [Pseudo...        41   0.065
gi|17546232|ref|NP_519634.1|  PROBABLE RIBONUCLEASE HI PROTE...        41   0.067
gi|28870867|ref|NP_793486.1|  ribonuclease HI [Pseudomonas s...        41   0.071
gi|13474791|ref|NP_106361.1|  transposase [Mesorhizobium lot...        40   0.078
gi|26990835|ref|NP_746260.1|  ribonuclease HI [Pseudomonas p...        40   0.11
gi|22978683|gb|ZP_00024428.1|  hypothetical protein [Ralston...        39   0.17
gi|22977454|gb|ZP_00023257.1|  hypothetical protein [Ralston...        39   0.17
gi|2760874|gb|AAD10223.1|  ORFB [Xanthomonas campestris pv. ...        39   0.18
gi|22978013|gb|ZP_00023785.1|  hypothetical protein [Ralston...        39   0.18
gi|15805924|ref|NP_294623.1|  ribonuclease H [Deinococcus ra...        39   0.23
gi|15597012|ref|NP_250506.1|  ribonuclease H [Pseudomonas ae...        39   0.27
gi|24111645|ref|NP_706155.1|  RNase HI, degrades RNA of DNA-...        38   0.43
gi|23106231|ref|ZP_00092685.1|  hypothetical protein [Azotob...        38   0.45
gi|494541|pdb|1RBS|  Ribonuclease H (E.C.3.1.26.4) Mutant W...         38   0.47   S
gi|13787086|pdb|1G15|A  Chain A, Co-Crystal Of E. Coli Rnase...        38   0.53   S
gi|23061138|ref|ZP_00086001.1|  hypothetical protein [Pseudo...        38   0.54
gi|12084647|pdb|1F21|A  Chain A, Divalent Metal Cofactor Bin...        38   0.57   S
gi|20150576|pdb|1JXB|A  Chain A, I53a, A Point Mutant Of The...        37   0.64   S
gi|4378703|gb|AAD19632.1|  unknown [Xanthomonas campestris p...        37   0.64
gi|23041530|ref|ZP_00072980.1|  hypothetical protein [Tricho...        37   0.67
gi|494049|pdb|1GOA|  Ribonuclease H (E.C.3.1.26.4) Mutant W...         37   0.79   S
gi|15799890|ref|NP_285902.1|  RNase HI, degrades RNA of DNA-...        37   0.82   S
gi|443224|pdb|1RDA|  Ribonuclease H (E.C.3.1.26.4) Mutant W...         37   0.84   S
gi|27479697|gb|AAO17224.1|  RnhA [Photorhabdus luminescens]            37   0.90
gi|21910590|ref|NP_664858.1|  putative transposase [Streptoc...        37   0.92
gi|20150465|pdb|1JL1|A  Chain A, D10a E. Coli Ribonuclease Hi          37   0.92   S
gi|494540|pdb|1RBR|  Ribonuclease H (E.C.3.1.26.4) Mutant W...         37   0.95   S
gi|494543|pdb|1RBU|  Ribonuclease H (E.C.3.1.26.4) Mutant W...         37   1.0    S
gi|443089|pdb|1LAV|  Ribonuclease H (E.C.3.1.26.4) Mutant W...         37   1.1    S
gi|230303|pdb|1RNH|  Selenomethionyl Ribonuclease H (E.C.3....         37   1.1    S
gi|443090|pdb|1LAW|  Ribonuclease H (E.C.3.1.26.4) Mutant W...         37   1.1    S
```

(b)

```
□ >gi|22955850|gb|ZP_00003651.1|   hypothetical protein [Nitrosomonas europaea]
          Length = 320

 Score = 48.9 bits (115), Expect = 3e-04
 Identities = 44/134 (32%), Positives = 60/134 (44%), Gaps = 14/134 (10%)

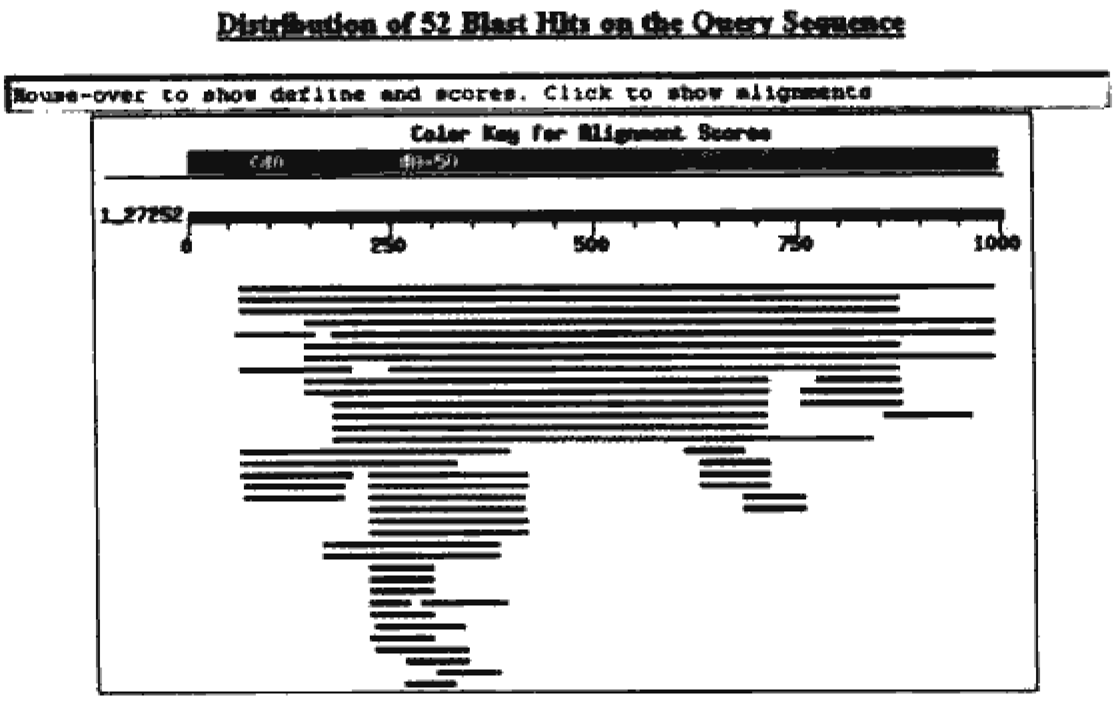
Query: 1216 EGKVIL-VAVHVASGYIEAEVIPAETGQETAYFLLKLAGRWPVK--TIHTDNGSNFTG-A 1271
             EGK+ L VA+    S +  AE++P       E A FL  L     P K  TI TDNG  FT
Sbjct: 146  EGKLYLFVAIDRTSKFAYAELLPKYGKNEAAQFLRNLIAAVPYKIHTILTDNGIQFTHRK 205

Query: 1272 TVRAA---------CWWAGIKQEFGIPYNPQSQGVVESMNKELKKIIGQVRDQAEHLKTAV 1323
             T R A         C    G +      P +P + G VE MN+ LK+     +        H +
Sbjct: 206  TDRHAFLHIFDRVCLENGTEHRLTQPNHPWTNGQVERMNRTLKEATVKRYHYENHQQLRE 265

Query: 1324 QMAVFI--HNFKRK 1335
              +  F+  +NF R+
Sbjct: 266  HLYSFLNAYNFARR 279
```

图 4-27 (a) 由一个使用HIV-1 pol蛋白进行的blastp搜索确定的细菌蛋白质，包括转座酶和核酸酶H蛋白；(b) 一个细菌蛋白质与一个HIV-1 pol查询序列的比对表明两个序列是同源的，在比对的延伸区域上具有显著的期望值和大约30%的氨基酸一致性

(a)



(b)

Sequences producing significant alignments:		Score (bits)	E Value
gi|5802821|gb|AAD51797.1|AF164614_1	(AF164614) Gag-Pro-Pol ...	239	2e-62
gi|5802814|gb|AAD51793.1|AF164611_1	(AF164611) Gag-Pro-Pol-...	238	2e-62
gi|5802819|gb|AAD51796.1|AF164613_1	(AF164613) Gag-Pro-Pol ...	238	2e-62
gi|13021822|gb|AAK11553.1|AF298587_1	(AF298587) polymerase ...	235	2e-61
gi|13021825|gb|AAK11554.1|AF298588_1	(AF298588) polymerase ...	235	2e-61
gi|130393|sp|P10266|POL1_HUMAN	RETROVIRUS-RELATED POL POLYP...	232	2e-60
gi|4456990|gb|AAD21097.1|	(AF074086) polymerase [Homo sapie...	232	2e-60
gi|1196429|gb|AAA88033.1|	(M14123) pol/env ORF (bases 3878-...	190	7e-48
gi|231033|pdb|4HVP|A	Chain A, HIV-1 Protease (HIV-1 PR) Com...	189	2e-47
gi|3600064|gb|AAC63290.1|	(AF080229) polymerase [Homo sapiens]	187	7e-47
gi|3600067|gb|AAC63291.1|	(AF080231) polymerase [Homo sapiens]	182	2e-45
gi|3600073|gb|AAC63294.1|	(AF080234) polymerase [Homo sapiens]	177	4e-44
gi|3600071|gb|AAC63293.1|	(AF080233) polymerase [Homo sapiens]	177	9e-44
gi|3600069|gb|AAC63292.1|	(AF080232) polymerase [Homo sapiens]	175	3e-43
gi|10504249|gb|AAG18012.1|	(AF248269) gag-pro-pol precursor...	145	2e-34
gi|5802810|gb|AAD51791.1|	(AF164609) Gag-Pro-Pol protein [H...	139	1e-32
gi|5802824|gb|AAD51799.1|AF164615_1	(AF164615) Gag-Pro-Pol ...	125	2e-28
gi|2257764|gb|AAB63113.1|	(U87590) polymerase [Homo sapiens]	118	3e-26
gi|2257766|gb|AAB63114.1|	(U87591) polymerase [Homo sapiens]	117	7e-26
gi|2257762|gb|AAB63112.1|	(U87589) polymerase [Homo sapiens]	111	4e-24
gi|2257773|gb|AAB63117.1|	(U87595) polymerase [Homo sapiens]	111	6e-24
gi|2257770|gb|AAB63116.1|	(U87593) polymerase [Homo sapiens]	107	7e-23
gi|2257768|gb|AAB63115.1|	(U87592) polymerase [Homo sapiens]	105	3e-22
gi|14760974|ref|XP_034809.1|	similar to putative gag-pro-po...	77	2e-13

图 4-28 (a) 使用 HIV-1 pol 蛋白进行 blastp 搜索寻找其在人类蛋白中的匹配的图形化输出结果；(b) 图中显示了 pol 蛋白在人类蛋白中的匹配要比其在细菌中的匹配多很多

表达，将产生 RNA 转录本，可用来自 cDNA 文库的 ESTs 来表现其特征。用病毒 pol 蛋白来检索人类 ESTs，有必要使用带有 tblastn 算法的 translating BLAST 站点，数据库要固定为 EST（图 4-29）。有成百上千的活跃转录的人类转录子被预测编码与病毒 pol 同源的蛋白，在第 6 章中将看到怎样评估这些 ESTs，确定它们在体内哪一部分或是在发育的哪个阶段表达。一些研究者提出像精神分裂症这样的精神病与体内过高水平的内源性逆转录病毒基因表达有关（Karlsson 等，2001）。

这些与 HIV-1 pol 同源的人类 ESTs 会不会和其他病毒 pol 蛋白更相近？要回答这个问题，选择与 HIV-1 pol 相关的一种人类 ESTs（图 4-30；我们选择 AI636743）。输入这个 EST 序列，将生物体限制为病毒，进行加强计算的 tblastn 检索。目前，这样的检索会鉴定出很多内源性的人类逆转录病毒序列（其 DNA 序列是人类基因组的一部分），在 GenBank 中这些序列被错误地归为病毒那一部分。与这个人类 EST 最相近的一个蛋白是一种可感染绵羊的病毒（序列号 AF105520）。从对 HIV 的 BLAST 检索开始，通过进一步一系列的检索探求了 HIV-pol 的生物学意义。

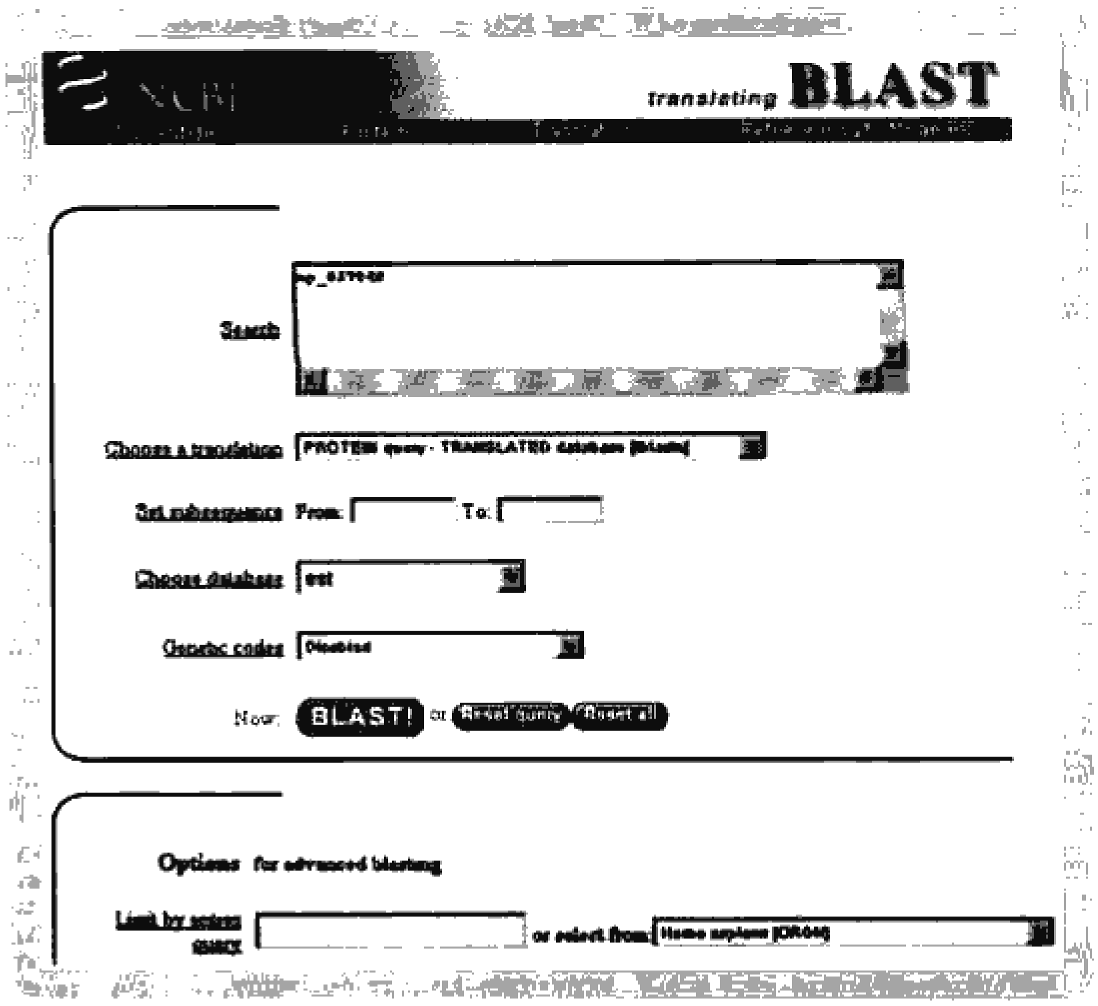


图 4-29 人类中表达的转录基因是否编码与HIV-1 pol蛋白同源的蛋白质？为回答这个问题，选择对一个人类EST数据库进行tblastn搜索并点击"BLAST!"按钮

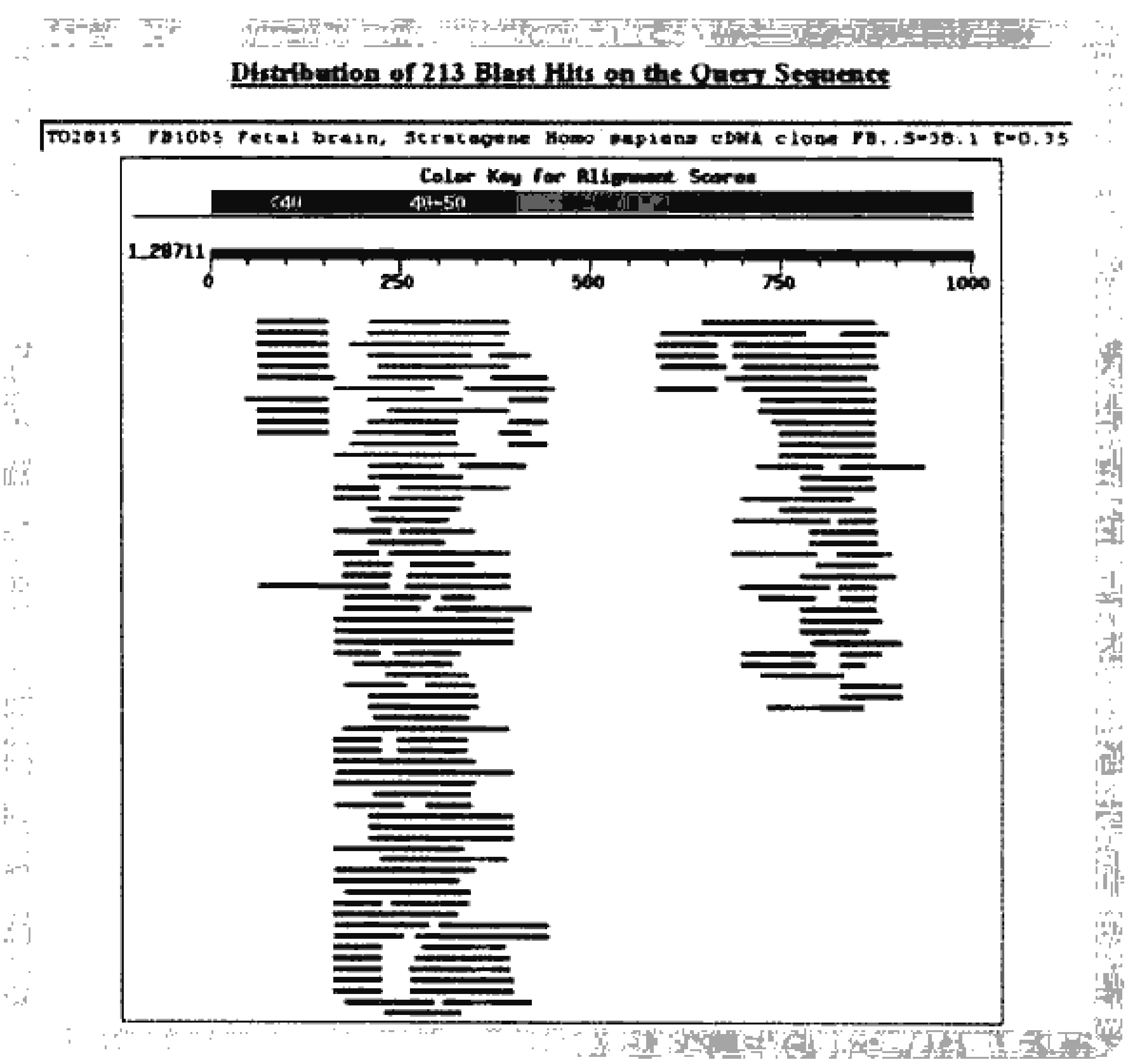


图 4-30 使用病毒pol蛋白对一个人类EST数据库进行tblastn搜索的结果，图中可以看到很多人类基因都被活跃地转录来产生被预测制造与HIV-1 pol蛋白同源的蛋白质的mRNA

4.6 BLAST 检索脂质运载蛋白：改变打分矩阵的作用

打分矩阵可能会对 BLAST 检索结果造成极大的影响，就像对比对对列那样（第 3 章）。我们用"human RBP4"作为 query 进行限制在人类蛋白中的 blastp nr 检索（图 4-18）。默认的 scoring matrice 是 BLOSUM 62，如果使用适合关系较近蛋白的 PAM 矩阵，将得到更少的结果。像载脂蛋白 D 这样的匹配项的期望值是 10^{-4} 而不是 10^{-8} [图 4-31(a)]。如果打分矩阵用 PAM 70，将得到额外的脂质运载蛋白的结果 [图 4-31(b)]。用大鼠 OBP 作

(a)

```
                                                                   Score    E
Sequences producing significant alignments:                        (bits) Value

gi|20141667|sp|P02753|RETB_HUMAN  Plasma retinol-binding pro...    578   e-165
gi|5803139|ref|NP_006735.1|  retinol-binding protein 4, plas...    576   e-164
gi|230284|pdb|1RBP|   Retinol Binding Protein >gi|493897|pdb...    574   e-164
gi|4558179|pdb|1QAB|E  Chain E, The Structure Of Human Retin...    554   e-158
gi|7770173|gb|AAF69622.1|AF119917_30  PRO2222 [Homo sapiens]       498   e-141
gi|296672|emb|CAA26553.1|  RBP [Homo sapiens]                      313   3e-85
gi|5419892|emb|CAB46489.1|  RBP (aa 101-172) [Homo sapiens]        226   4e-59
gi|2895204|gb|AAC02945.1|  mutant retinol binding protein [H...    138   2e-32
gi|2895206|gb|AAC02946.1|  mutant retinol binding protein [H...    109   6e-24
gi|619383|gb|AAB32200.1|  apolipoprotein D, apoD (human, pla...     45   2e-04
gi|1246096|gb|AAB35919.1|  apolipoprotein D; apoD [Homo sapi...     44   3e-04
gi|4502163|ref|NP_001638.1|  apolipoprotein D precursor [Hom...     44   3e-04
gi|4506251|ref|NP_000945.1|  prostaglandin D2 synthase 21kDa...     32   1.0
gi|404390|gb|AAB27607.1|  beta-trace protein, prostaglandin ...     32   1.7
gi|190444|gb|AAA36494.1|  prostaglandin D synthase precursor...     32   1.7
gi|27480905|ref|XP_209977.1|  similar to CALII [Gallus gallu...     30   4.2
```

(b)

```
                                                                   Score    E
Sequences producing significant alignments:                        (bits) Value

gi|20141667|sp|P02753|RETB_HUMAN  Plasma retinol-binding pro...    498   e-141
gi|5803139|ref|NP_006735.1|  retinol-binding protein 4, plas...    496   e-140
gi|230284|pdb|1RBP|   Retinol Binding Protein >gi|493897|pdb...    495   e-140
gi|4558179|pdb|1QAB|E  Chain E, The Structure Of Human Retin...    480   e-135
gi|7770173|gb|AAF69622.1|AF119917_30  PRO2222 [Homo sapiens]       430   e-120
gi|296672|emb|CAA26553.1|  RBP [Homo sapiens]                      269   5e-72
gi|5419892|emb|CAB46489.1|  RBP (aa 101-172) [Homo sapiens]        196   3e-50
gi|2895204|gb|AAC02945.1|  mutant retinol binding protein [H...    119   6e-27
gi|2895206|gb|AAC02946.1|  mutant retinol binding protein [H...     97   4e-20
gi|619383|gb|AAB32200.1|  apolipoprotein D, apoD (human, pla...     61   3e-09
gi|4502163|ref|NP_001638.1|  apolipoprotein D precursor [Hom...     59   6e-09
gi|1246096|gb|AAB35919.1|  apolipoprotein D; apoD [Homo sapi...     50   4e-06
gi|223373|prf||0801163A  complex-forming glycoprotein HC            38   0.019
gi|4502067|ref|NP_001624.1|  alpha-1-microglobulin/bikunin p...     35   0.16
gi|4506251|ref|NP_000945.1|  prostaglandin D2 synthase 21kDa...     31   2.1
gi|404390|gb|AAB27607.1|  beta-trace protein, prostaglandin ...     30   3.1
gi|190444|gb|AAA36494.1|  prostaglandin D synthase precursor...     30   3.1
gi|27480905|ref|XP_209977.1|  similar to CALII [Gallus gallu...     30   3.6
gi|6841194|gb|AAF28950.1|AF161390_1  HSPC272 [Homo sapiens]         29   5.7
gi|2224565|dbj|BAA20771.1|  KIAA0312 [Homo sapiens]                 29   5.9
gi|22090626|dbj|BAC06833.1|  HECT domain protein LASU1 [Homo...     29   6.3
gi|22538476|ref|NP_113584.2|  upstream regulatory element bi...     29   6.3
gi|21730336|pdb|1IW2|A  Chain A, X-Ray Structure Of Human Co...     29   6.6
```

图 4-31 使用 RBP4 作为查询序列的 blastp nr（人类）搜索的结果，这里分别使用了（a）PAM30 打分矩阵或（b）PAM70 矩阵。将（a）中的结果与使用默认的 BLOSUM62 矩阵（图 4-18）得到的结果进行比较：亲缘关系较远的蛋白（如载脂蛋白 D）当使用 PAM30 矩阵时会得到很差的打分。（b）中的 PAM70 矩阵被设计来从数据库中找到更多的亲缘关系较远的匹配，这里得到的一些蛋白在使用 BLOSUM62 或 PAM30 矩阵时都是没有出现的

为 query 对人类蛋白的检索只得到 5 个匹配项（其中只有 3 个是可信的脂质运载蛋白）（图 4-32）。将矩阵变回 BLOSSUM 45，此矩阵倾向于检测到氨基酸较少相同的蛋白。这里得到 12 个匹配项，包括四个脂质运载蛋白。

```
                                                               Score    E
Sequences producing significant alignments:                    (bits) Value

gi|17433367|ref|XP_066998.1|  similar to odorant-binding pro...   36   0.024
gi|7657407|ref|NP_055396.1|  odorant-binding protein 2B [Hom...   35   0.042
gi|14585875|gb|AAK67638.1|  proteoglycan link protein [Homo ...   33   0.15
gi|20177973|sp|Q8NX39|NUPL_HUMAN  Putative NUP-like lipocali...   31   0.79
gi|17483447|ref|XP_065151.1|  similar to cytochrome P450, su...   30   1.6

                                  Alignments

>gi|17433367|ref|XP_066998.1|  similar to odorant-binding protein [Elephas maximus] [Homo sapiens]
          Length = 103

 Score = 36.2 bits (82), Expect = 0.024
 Identities = 18/58 (31%), Positives = 32/58 (55%)

Query: 27   EVNGDWRTLYIVADNVEKVAEGGSLRAYFQHMECGDECQELKIIFNVKLDSECQTHTV 84
             +V GD   T+++  D+VEK+ E G + +T   +EC +  + L   F +K +S   T +
Sbjct: 46   KVAGDRYTIFMAVDDVEKIQENGPITSYIHWIECYNGGETLIFTFYMKDNSAEHTDDI 103

>gi|7657407|ref|NP_055396.1|  odorant-binding protein 2B [Homo sapiens]
 gi|20139234|sp|Q9NPH6|OBPB_HUMAN  Odorant-binding protein 2b precursor (OBPIIb)
 gi|6900075|emb|CAB71322.1|  odorant binding protein b [Homo sapiens]
 gi|6900077|emb|CAB71323.1|  putative odorant-binding protein b-a [Homo sapiens]
          Length = 170

 Score = 35.4 bits (80), Expect = 0.042
 Identities = 42/185 (22%), Positives = 71/185 (38%), Gaps = 35/185 (18%)

Query: 2    VKFLLIVLALGVSCAHHENLDISPSEVNGDWRTLYIVADN--------VEKVAE-------- 47
             +K L + + LG++ A    +    ++ G W    +V D            KV+
Sbjct: 1    MKTLFLGVTLGLAAAL--SFTLEEEDITGTWYVKAMVVDKDFPEDRRPRKVSPVKVTALG 58

Query: 48   GGSLRAYFQHMECGDECQELKIIFNVKLDSECQTHTVVGQKHEDGRYTTDYSGRN--YFH 105
             GG L A F  H   D C + KI+                +  E G+Y+  Y GR   Y
Sbjct: 59   GGKLEATFTFMR-EDRCIQKKILMR---------------KTEEPGKYSA-YGGRKLMYLQ 102

Query: 106  VLKKTDDIIFFHNVNVDESGRRQCDLVAGKREDLNKAQKQELRKLAEEYNIPNENTQHLV 165
              L + D  IF+   +    G       + G+  D N+   +E  +KL  +   + E+      +
Sbjct: 103  ELPRRDHYIFYCK-DQHHGGLLHMGKLVGRNSDTNREALEEFKKLVQRKGLSEEDIFTPL 161

Query: 166  PTDTC 170
              T +C
Sbjct: 162  QTGSC 166
```

图 4-32 使用小鼠的气味分子结合蛋白作为查询项在人类蛋白质中进行的 blastp 搜索的结果

4.7 展望

BLAST 检索已经成为一种不可缺少的工具，被用于分析某个 DNA 或蛋白与公用数据库中成百万序列之间的关系。所有的数据库检索工具都必须解决灵敏度（将假阴性结果最小化的能力）和选择性（使假阳性结果最小化的能力）以及时间的问题。随着近些年来公用数据库规模的膨胀，BLAST 工具已经成为一种筛选数据库的快速、可靠的途径。

4.8 常见问题

在 BLAST 检索中要避免几种常见的缺陷。在 BLAST 初学用户中常见的错误是选择了错误的数据库来搜索蛋白或 DNA 序列。理解基本的 BLAST 算法是必要的。图 4-3 对这些概念做出了总结。

4.9 网络资源

BLAST 检索的主要站点是国家生物技术信息中心（http://www.ncbi.nlm.nih.gov/blast)。这个站点提供了一些主要程序的链接（blastn，blastp，blastx，tblastn，tblastx)。第 5 章将讨论一些其他的特殊 BLAST 站点，作为 BLAST 检索的另一种选择。我们可以检索其他的数据库，例如 Pfam 和 Ensembl，第 5 章也将讨论这些数据库检索工具。

一个重要的网络资源是 NCBI 站点提供的 BLAST 指南、课程和参考文献。

4.10 讨论题

1. 为什么没有人提供“Basic Global Alignment Search Tool (BGAST)”来补充 BLAST？BGAST 会成为一种有用的工具吗？创立它可能遇到的计算上的困难是什么？

2. 你认为 1，0.05 或者 10^{-3} 是有意义的期望值吗？答案是否依赖于你所做检索的种类？

4.11 问题

1. NCBI BLAST 网页提供了进行特殊检索的指南（http://www.ncbi.nlm.nih.gov/BLAST/；见图 4-33)。首先猜测期望值、字符大小以及 low-complexity filtering 的推荐值

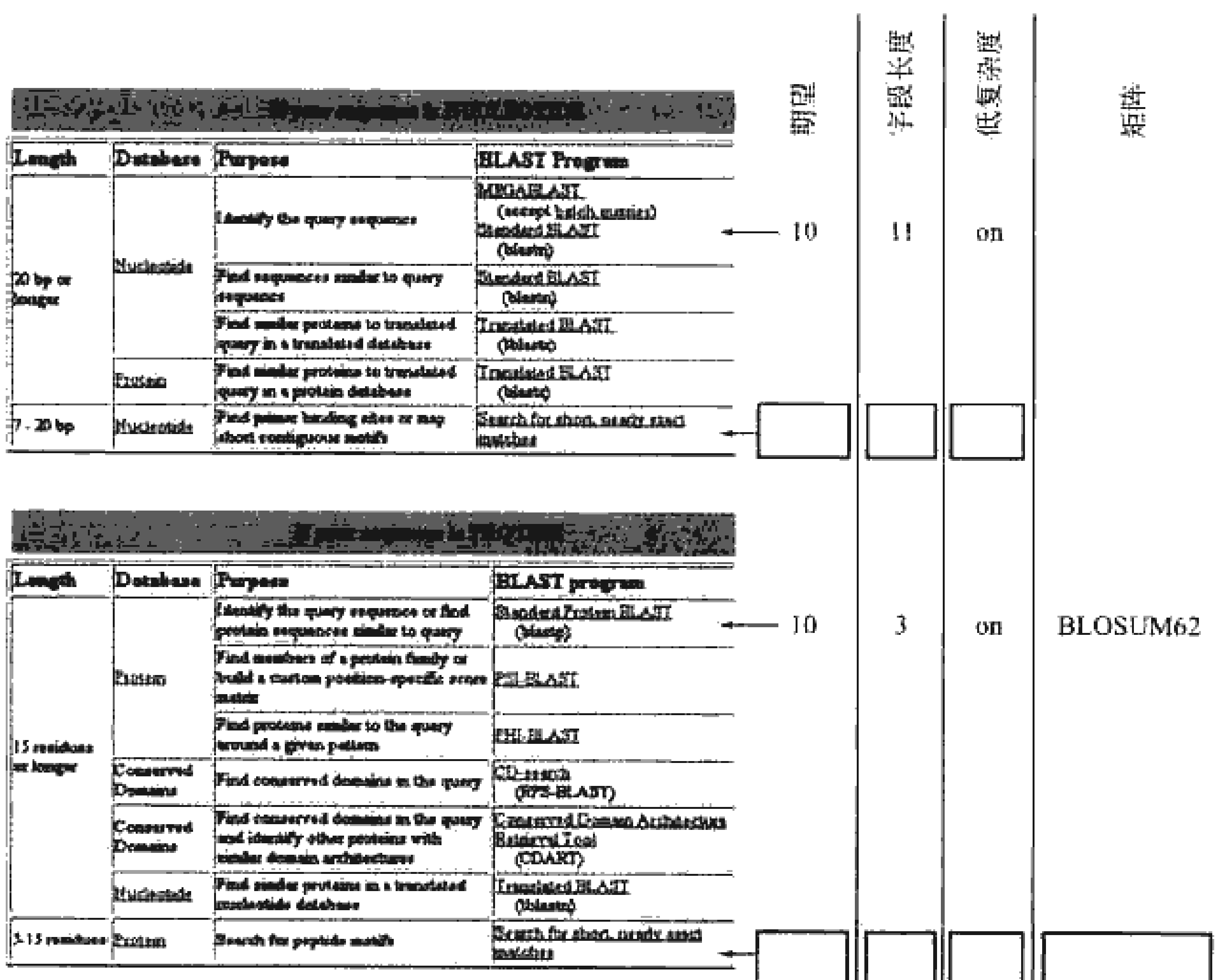


图 4-33 NCBI BLAST 站点上包含的对于使用核酸或蛋白质序列进行各种搜索的向导。对于短的查询序列，要改变期望值和字段长度；对于蛋白质搜索：建议使用的打分矩阵也要改变。猜测一下这些都要改成什么值，然后点击 BLAST 向导（http://www.ncbi.nlm.nih.gov/BLAST）核对一下

是什么，然后到 NCBI 站点点击短核苷酸核和小蛋白 queries 的超链接，检查你的回答是否正确。

2. 为什么像 BLAST 这样的数据库程序必须在灵敏性和选择性之间达到平衡？blastp 算法是怎样解决这个问题的？

3. 蛋白检索包含的信息通常比 DNA 检索大。用 RBP4（NP _ 006735）做 blastp 检索，将输出限制在 Arthropoda（昆虫）。接着，用 RBP4 的核苷酸序列（NM _ 006744，选择与 DNA 编码区相对应的核苷酸）做 blastn 检索。哪一个检索能提供更多的信息？在各个检索中，有多少数据库匹配项的 E 值低于 1.0？

4. 人类中最大的基因家族被认为是嗅觉受体家族。做 BLAST 检索，估计这个家族的大小。提示：作为一种策略，先通过 locuslink 进入“嗅觉受体”，将生物体限制为人类（*homo sapiens*）。得到 1100 个以上的条目，但这并没有告诉它们之间是否相互关联。选择其中一个序列号，进行将生物体限制为人类的 blastp 检索。

5. 对于第 4 题中的检索，如果选择了一个打分矩阵更适合查找关系较远的蛋白，结果将怎样？

6. 气味分子结合蛋白（OBP）在大鼠中被发现，并作为牛的一种脂质运载蛋白，可选择性结合气味分子。找到这种大鼠蛋白的序列号（提示：尝试 Entrez 蛋白）检索最相近的人类同源物，它是什么？

7. HIV-2 和猿免疫缺陷性病毒中哪一个 pol 蛋白同 HIV-1 的蛋白更为接近？通过 blastp 程序回答这个问题。

8. “雪人（Iceman）”是一个生活在 5300 年前的人，1991 年他的尸体在意大利阿尔卑斯山被发现。在他的衣物上发现了一些真菌并对之进行了测序。与真菌 DNA 最为相近的现代物种是什么？

9. 如果进行了一个 BLAST 检索，E 值是 10^{-29}。这个 E 值的意义是什么？E 值的大小依赖于哪些参数？

4.12 自测题

1. 如有一小段 DNA 序列。基本上它能编码多少种蛋白？

(a) 1；

(b) 2；

(c) 3；

(d) 6

2. 有一段 DNA 序列，如想知道在主要的蛋白数据库（“nr”，去冗余数据库）中哪一个与该 DNA 编码的蛋白最相近，你将选择用哪一个程序？

(a) blastn；

(b) blastp；

(c) blastx；

(d) tblastn；

(e) tblastx

3. BLAST 检索的那一种输出估计了假阳性的数目？

(a) E 值；

(b) Bit score；

(c) Percent identity；

(d) Percent positives

4. 将 BLAST 检索程序与对它们的正确

描述连线

blastp (a) 对于核苷酸序列数据库进行的核苷酸 query

blastn (b) 对于翻译后的核苷酸序列数据库进行的蛋白 query

blastx (c) 对于蛋白数据库进行的翻译后核苷酸的 query

tblastn (d) 对于蛋白数据库进行的蛋白 query

tblastx (e) 对于翻译后的核苷酸序列数据库进行的翻译后核苷酸的 query

5. 将下面哪个 BLAST 参数改变后会得到更少的检索结果?

(a) 关闭 low-complexity filter;

(b) 将期望值从 1 变为 10;

(c) 提高极限值;

(d) 将打分矩阵从 PAM30 改为 PAM70

6. 可以用任意的 Entrez 条件限制 BLAST 检索，例如，可以通过研究者的名字对结果做出限制。这句话对否?

(a) 正确;

(b) 错误

7. 极值分布

(a) 描述了对数据库的 query 的 scores 的分布;

(b) 比正态分布的总面积大;

(c) 对称;

(d) 形状可用两个常数来描述，即 μ (平均值) 和 λ (衰减常数)

8. 当 BLAST 检索的 E 值减小时

(a) K 值也减小;

(b) score 变大;

(c) 概率 p 值变大;

(d) 极值分布偏斜度减小

9. BLAST 算法汇编了 1 个代表 3 个氨基酸的“字符”的清单(蛋白质检索)。$T \geq$ 极限值的字符被定义为

(a) hits，用来浏览数据库以找到准确吻合的匹配项，随后可被扩充;

(b) hits，用来浏览数据库以找到准确或部分吻合的匹配项，随后可被扩充;

(c) 并列在一起的“hits”;

(d) 作为 raw scores 的“hits”

10. 标准化的 BLAST score (也称为 bit scores)

(a) 是没有单位的;

(b) 与使用的打分矩阵无关;

(c) 可在不同的 BLAST 检索之间比较，即使使用了不同的打分矩阵;

(d) 可在不同的 BLAST 检索之间比较，但前提是使用相同的打分矩阵

4.13 推荐读物

BLAST 检索第一次被提出是在 Stephen Altschul，David Lipman 及同事的一篇经典文献(1990)中。这篇论文描述了 BLAST 检索的理论基础以及一些基本问题，例如灵敏度(正确度)和速度。晚些时候有对 BLAST 算法的重要修饰，包括间隔 BLAST 的引入(Altschul 等，1997)。这篇论文包括了对于位置特异性打分矩阵的讨论，第 5 章将考虑此问题。

William Pearson (1996) 在一篇题为“有效蛋白序列比对”的文章中精彩地描述了数据库检索。Altschul 等 (1994) 发表了一篇被高度评价的文章——“检索分子序列数据库中的问题”。瑞士生物信息学研究中心的 Marco Pagni 和 C. Victor Jongeneel (2001) 对序列比对统计学进行了技术上的总结。这篇文章包括极值分布、随意序列的使用、带有或不带有间隔的局部比对、BLAST 统计学。Stephen Altschul 和 arren Gish 也有一篇精彩的有关比对统计学的综述 (1996)。

参 考 文 献

Altschul, S. F., Boguski, M. S., Gish, W., & Wootton, J. C. Issues in searching molecular sequence databases. *Nat. Genet.* **6**, 119-129 (1994).

Altschul, S. F., and Gish, W. Local alignment statistics. *Methods Enzymol.* **266**, 460-480 (1996).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410. (1990).

Altschul, S. F., et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).

Brenner, S. E. Practical database searching. *Bioinformatics: A Trends Guide* **1998**, 9-12 (1998).

Chambaud, I., Heilig, R., Ferris, S., Barbe, V., Samson, D., Galisson, F., Moszer, I., Dybvig, K., Wroblewski, H., Viari, A., Rocha, E. P., Blanchard, A. The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acids Res.* **29**, 2145-2153 (2001).

Dixon, R. A., et al. Cloning of the gene and cDNA for mammalian beta-adrenergic receptor and homology with rhodopsin. *Nature* **321**, 75-79 (1986).

Downward, J., et al. Close similarity of epidermal growth factor receptor and *v-erb-B* oncogene protein sequences. *Nature* **307**, 521-527 (1984).

Ermolaeva, M. D., White, O., and Salzberg, S. L. Prediction of operons in microbial genomes. *Nucleic Acids Res.* **29**, 1216-1221 (2001).

Ferretti, J. J., et al. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc. Natl. Acad. Sci. USA* **98**, 4658-4663 (2001).

Gish, W., and States, D. J. Identification of protein coding regions by database similarity search. *Nat. Genet.* **3**, 266-272 (1993).

Gotoh, O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.* **264**, 823-838 (1996).

Gumbel, E. J. *Statistics of Extremes*. Columbia University Press, New York, 1958.

International Human Genome sequencing consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

Karlin, S., and Altschul, S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* **87**, 2264-2268 (1990).

Karlsson, H., et al. Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia. *Proc. Natl. Acad. Sci. USA* **98**, 4634-4639 (2001).

Needleman, S. B., and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443-453 (1970).

Pagni, M., and Jongeneel, C. V. Making sense of score statistics for sequence alignments. *Brief. Bioinform.* **2**, 51-67 (2001).

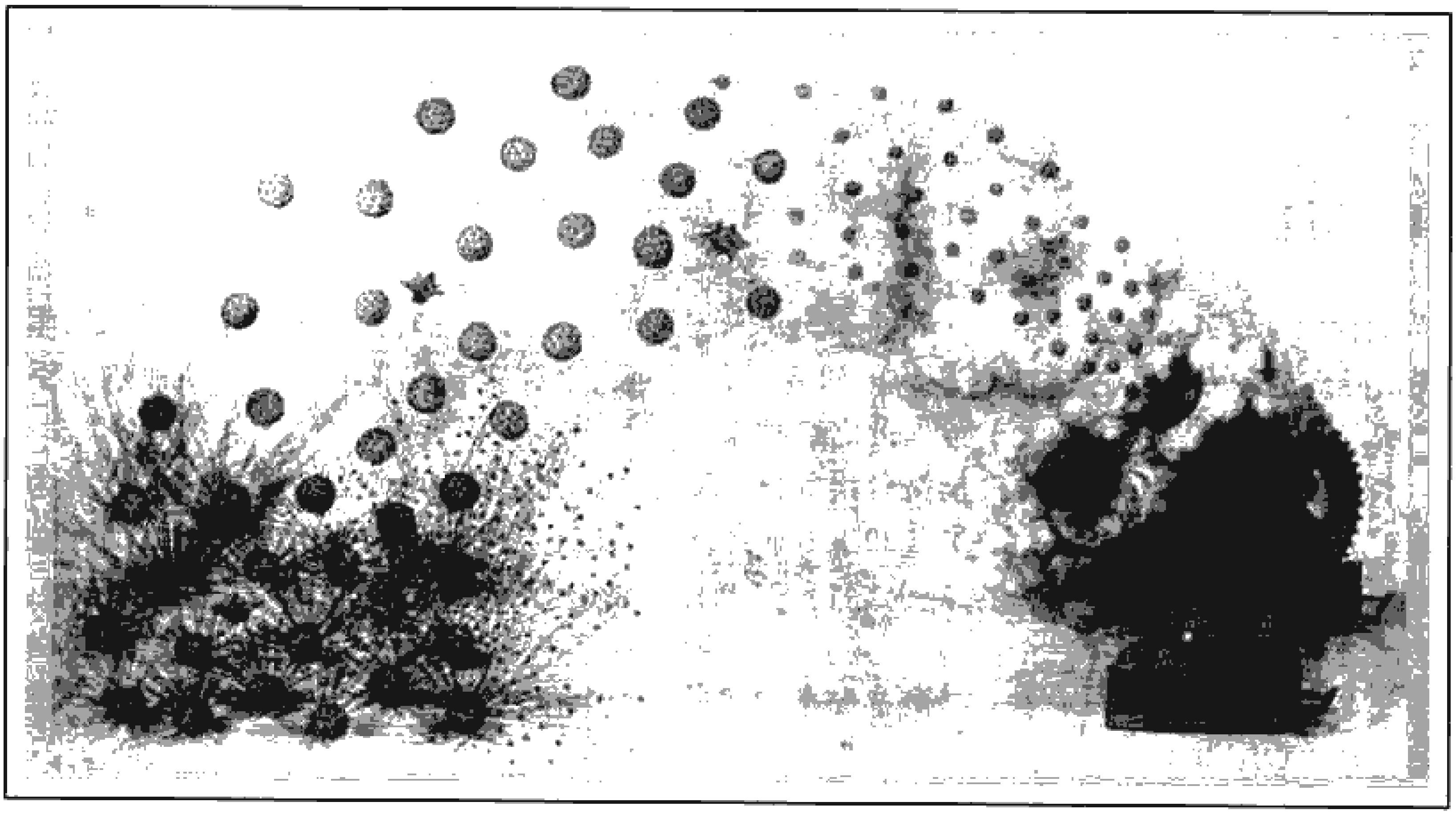
Park, J., et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201-1210 (1998).

Pearson, W. R. Effective protein sequence comparison. *Methods Enzymol.* **266**, 227-258 (1996).

Smith, T. F., and Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197 (1981).

Tettelin, H., et al. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* **293**, 498-506 (2001).

Wootton, J. C., and Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554-571 (1996).



达·芬奇设计了具有爆炸性外壳的砂浆（引自 Codex 集，第 33 卷）。图中所有的球都被小洞填充以获得最大的爆破力。这幅画始作于 1845 年

第5章

高级比对搜索

5.1 引言

在第3章和第4章中介绍了序列间的比拼和比对搜索。通过比对搜索可以从一个数据库中找出感兴趣的一些蛋白质或者基因，然后进行研究。比对搜索是非常有用的一种研究手段，所以在本章中继续介绍比对搜索的高级技术。

本章首先简要介绍各种专门而又常用的比对资源。为什么需要专门的比对网站呢？因为在现实的研究中我们常常会遇到一些非常专业的问题，而各种不同的专业化比对网站的存在是生物信息学研究中必不可少的条件。如我们常常需要用比对搜索来解决下面的一些问题：某个特定生物中是否存在一个或多个与我们感兴趣的基因相关的基因？我们拿到的一个蛋白质都含有哪些结构域？我们感兴趣的基因是否在微生物中存在同源物？本章还将介绍一下怎么用比对搜索来发现一些新的基因。

5.2 专门的 Blast 比对网站

到目前为止，已经使用了两种比对资源，它们都来自 NCBI 的网站：比对2序列和标准5比对程序。本章将介绍其他专门的比对（或者是比对相关的网站）程序。这些专门的比对网站可以简要地分为以下几类：基于某些具体生物的比对网站，可以搜索某些特定生物分子的比对网站，以及专门提供一些具体算法的数据库。这些专门的算法可以很大程度地提高比对搜索的敏感性。

5.2.1 基于某些具体生物的比对网站

我们已经看到了在 NCBI 网站上标准的比对搜索得到的结果可以限定于某一特定的生物。现在已经存在很多其他专门用来储存关于某种生物体内所有分子序列结构的数据库（表5-1）。这些数据库往往含有一些还未完全测序因而未被 GenBank 收录的序列结构信息。因此，可以通过搜索这些数据库获得一些非常有用的信息。例如，当你拿到了一个蛋白质或者 DNA 的序列，而且在 NCBI 中进行标准的比对搜索时并未获得一个匹配，那么，你可以搜索这些专门的比对数据库，它们往往能够提供更大范围内的搜索。另外，正如下文所述，这些专门的比对数据库往往以某种特殊的格式来显示搜索结果而且提供一些专门的搜索算法。

表 5-1 特定物种的 BIAST 和 FASTA 网站

物种	描述	网址
细菌	蓝细菌	http://www.kazusa.or.jp/cyano/search.html
天蓝色链霉菌(*Streptomyces coelicolor*)	桑格研究中心	http://www.sanger.ac.uk/Projects/S_coelicolor/blast_server.shtml
网状细菌(*Dictyostelium discoideum*)	Dicty 比对	http://glamdring.ucsd.edu/others/dsmith/dictydb.html
	D. discoideum 基因组计划	http://genome.imb-jena.de/dictyostelium/BlastDictyForm.html
真菌		
栗酒裂殖酵母(*Schizosaccharomyces pombe*)	桑格研究中心	http://www.sanger.ac.uk/Projects/S_pombe/blast_server.shtml
曲霉菌(*Aspergillus fumigatus*)	桑格研究中心	http://www.sanger.ac.uk/Projects/A_fumigatus/
白色念珠菌(*Candida albicans*)	桑格研究中心	http://www.sanger.ac.uk/Projects/C_albicans/Urlbblast_server.shtml
粗糙脉孢菌(*Neurospora crassa*)	Whitehead 研究中心	http://www-genome.wi.mit.edu/annotation/fungi/neurospora/
Pneumocystis carinii	桑格研究中心	http://www.sanger.ac.uk/Projects/P_carinii/blast_server.shtml
酿酒酵母(*Saccharomyces cerevisiae*)	MIPS	http://mips.gsf.de/cgi-bin/blast/blast_page? genus=yeast
白色杆菌	能源部和基因组联合研究中心	http://www.jgi.doe.gov/programs/whiterot/whiterot_mainpage.html
昆虫		
果蝇(*Drosophila melanogaster*)	果蝇比对	http://www.fruitfly.org/blast/
蚊子	一个源于蚊子基因组网络服务器的 NCBI 比对(http://klab.agsci.colostate.edu/)	http://mosquito.colostate.edu/cgi-bin/ace/blast/MsqDB
冈比亚按蚊(*Anopheles gambiae*)	AnoDB	http://konops.imbb.forth.gr/cgi-bin/blast2.pl
微生物基因组	可以比对完整的和不完整的基因组	http://www.ncbi.nlm.nih.gov/BLAST/
	可以比对完整的和不完整的基因组	http://www.tigr.org
	TIGR 上丰富的微生物资源	http://tigrblast.tigr.org/cmr-blast/
	数十种特定生物的比对搜索	http://www.sanger.ac.uk/Projects/Microbes/
寄生虫		
变种	位于 EBI	http://www.ebi.ac.uk/blast2/parasites.html
	巴西的 Oswaldo Cruz 研究中心	http://www.dbbm.fiocruz.br/genome/parasite-genome/parasite_blast_server.html
蚜虫(*Phytophthora*)	一种植物害虫(http://www.ncgr.org)	http://www.ncgr.org/pgc/blast/blast.html
植物:拟南芥(*Arabidopsis thaliana*)	Thale cress	http://genome-www2.stanford.edu/cgi-bin//AtDB/nph-blast2atdb
原生动物		
牛巴贝虫(*Babesia bovis*)	桑格研究中心	http://www.sanger.ac.uk/Projects/B_bovis/blast_server.shtml
贾第虫(*Giardia lamblia*)	海洋生物实验室	http://www.mbl.edu/Giardia/blast.html
疟原虫的一种(*Plasmodium chabaudi*)	桑格研究中心	http://www.sanger.ac.uk/Projects/P_chabaudi/blast_server.shtml
疟原虫的一种(*Plasmodium falciparum*)	这就是引发疟疾的生物	http://www.ncbi.nlm.nih.gov/BLAST/
	Plasmodium 基因组资源	http://plasmodb.org/
	桑格研究中心	http://www.sanger.ac.uk/Projects/P_falciparum/blast_server.shtml
方盲珊瑚(*Tetrapygus niger*)	海胆比对	http://sea-urchin.caltech.edu/genome/sequence/blast.py
其他物种	TIGR 比对	http://www.tigr.org
	桑格研究中心的 OmniBLAST 服务器	http://www.sanger.ac.uk/DataSearch/omniblast.shtml
	Gene 2EST 比对服务器	http://woody.embl-heidelberg.de/gene2est/
	oral 比对服务器	http://199.93.248.241/blast/blast.html

续表

物种	描述	网址
脊椎动物		
玻璃海鞘(*Ciona intestinalis*)	海鞘	http://www.jgi.doe.gov/programs/ciona/ciona_mainpage.html
Danio rerio	桑格研究中心的斑马鱼资源	http://www.sanger.ac.uk/Projects/D_rerio/blast_server.shtml
河豚(*Fugu*)	包括斑马鱼	http://fugu.hgmp.mrc.ac.uk/blast/
	能源部和基因组联合研究中心	http://www.jgi.doe.gov/programs/fugu/fugu_mainpage.html
一种禽类(*Gallus gallus*)	来自 ChickMap(http://www.ri.bbsrc.ac.uk/chickmap/ChickMap HomePage.html)	http://www.ri.bbsrc.ac.uk/cgi-bin/est-blast/blast.pl
人(*Homo sapiens*)	允许针对特定染色体的比对搜索	http://www.ncbi.nlm.nih.gov/BLAST/
	主人类基因组网站	http://www.Ensembl.org
家鼠(*Mus musculus*)	老鼠特定染色体的比对搜索	http://www.ncbi.nlm.nih.gov/genome/seq/MmBlast.html
	杰克逊实验室的老鼠基因组信息中心	http://mouseblast.informatics.jax.org/prototype/
	桑格研究中心和 Ensembl	http://mouse.Ensembl.org/
	桑格研究中心	http://mouse.Ensembl.org/perl/blastview
	国家老龄研究中心	http://lgsun.grc.nia.nih.gov/cDNA/cDNA.html
蟾蜍(*Xenopus*)	包括来自 *X. laevis* 和 *X. tropicalis* 的数据	http://www.sanger.ac.uk/Projects/X_tropicalis/blast_server.shtml
蠕虫		
马来丝虫(*Brugia malayi*)	桑格研究中心	http://www.sanger.ac.uk/Projects/B_malayi/blast_server.shtml
	B. Malayi 比对服务器	http://www.dbbm.fiocruz.br/genome/parasite_genome/Brugia_blast_server.html
Caenorhabditis elegans	蠕虫比对，也包括其他物种；华盛顿大学的基因测序中心	http://genome.wustl.edu/gsc/Blast/client.pl
线虫(*C. elegans*)	桑格研究中心	http://www.sanger.ac.uk/Projects/C_elegans/blast_server.shtml
线虫(*C. elegans*)	史蒂夫·琼斯基因组计划 DNA/蛋白质数据库	http://www.bio.ic.ac.uk/fgn/blastform/elegansblast.html

注：这些网站提供一些含有未被 NCBI 收录的序列信息的数据序接口服务，有些站点使用 WUBLAST 2.0 比对算法。

5.2.1.1 Ensembl 比对

Ensembl 计划是由桑格研究所和欧洲生物信息中心联合发起的。Ensembl 网站提供了很多研究人类基因组和其他基因组的数据资源（见第 16 章和第 17 章）。Ensembl 比对服务器允许用户搜索它维护的数据库，其中包含了很多已经完全测定的序列。例如，输入文字形式的人 RBP4（序列号 NP_006735）的氨基酸序列，然后进行 tblastn 比对搜索（图 5-1）。结果显示了一个匹配（p 值是 3.8×10^{-81}；AL356214.00005）。点击这一链接可以打开一个网页，上面包含所在的染色体条带、邻近的基因、蛋白质及 DNA 数据库的链接（包括 NCBI 链接）、多态性、老鼠同源物以及表达数据等。这一结果还可以用图形的形式显示出匹配结果在染色体上的定位（图 5-2）。

Wellcome Trust Sanger Institute 的网站是 http://www.Sanger.ac.uk/。EBI 的网站是 http://www.ebi.ac.uk/。Ensemble 的人类 BLAST 服务器是 http://www.ensembl.org/perl/blastview，小鼠和其他动物物种的 Ensemble BLAST 服务器能在 http://www.ensembl.org/ 上找到。我们将会在第 12 章和第 17 章讨论"完成测序的"序列。

5.2.1.2 基因组研究所（TIGR）的比对搜索

TIGR 作为一个研究机构已经完整地测完了许多物种的基因组序列，而且 TIGR 的科学家们已经测出了另外一些物种的部分基因组序列，包括人类基因组、水稻基因组、土豆基因组、一些寄生虫

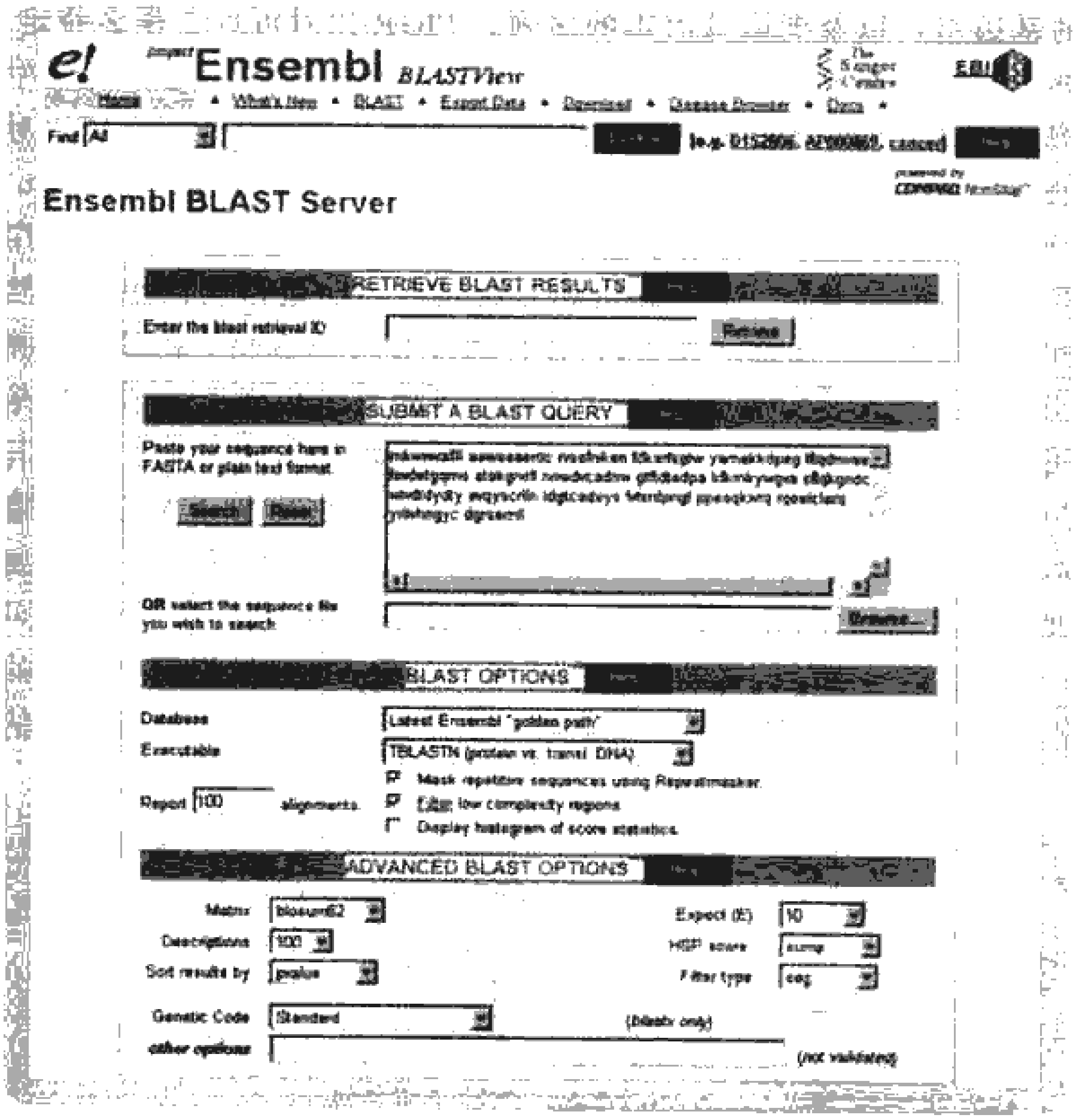


图 5-1 Ensembl BLAST 服务器（http://www.ensembl.org/perl/blastview）提供在人类基因组数据库中查询的服务

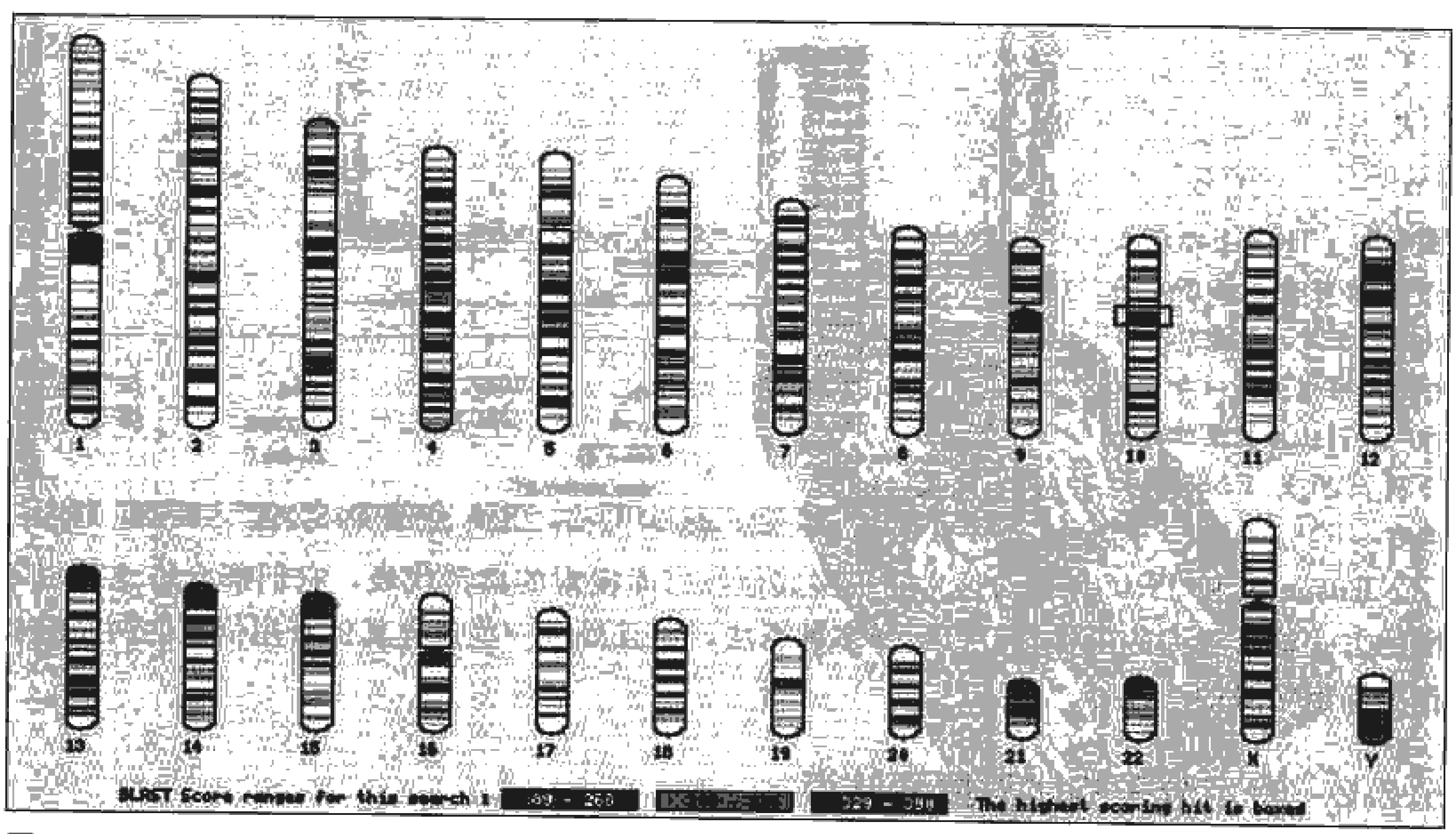


图 5-2 在 Ensemble 数据库中用 RBP4 蛋白质进行 BLAST 搜索后的结果。结果是以染色体的图像形式给出的，最好的匹配在 10 号染色体着丝粒附近的长臂上。其他染色体上较弱的匹配也很明显地标出来了

的基因组、一些微生物的基因组，以及老鼠基因组的序列。TIGR 比对可以在很多特定的数据库中搜索，而且这些搜索还可以让用户从 80 种替换矩阵中任意选择。进入 TIGR 网页，选择 TIGR 基因索引，然后选择比对搜索。在序列输入框内输入人 RBP4 蛋白质序列（该序列来自 Entrez 中的 NP _ 006735 对应的记录）（图 5-3）。注意序列号可能用不到，所以直接以 FASTA 格式复制然后粘贴蛋白质的序列。然后选择适当的算法（蛋白质序列搜索用 tblastn 算法），并且选择 EGO（真核生物基因同源物）。比对的结果以概率分数的高低顺序排列显示（注意该结果的显示用到了 p 值而不是 e 值）（图 5-4）。匹配以可能同源的或是可能与人类某种蛋白质同源的形式显示。点击其中的链接就可以看到该结果的 DNA 序列及与其对应的 EST 描述，它在生物体内的表达概况（包括该蛋白质在 cDNA 库中的空间区域定位，即该蛋白是从哪一个部位分离得到的），以及一个显示了其他物种中 RBP 同源物的表格。另外，该结果中还包含一个指向可能的同源物的链接（图 5-5）。这些可能的同源物以图形的形式显示 TIGR 数据库中已知的同源物，包括一个序列比对表（图 5-5 中显示了该表的部分内容以及该组内多序列比拼的结果）。

TIGR 的网址是 http://www.tigr.org，其中有一个服务器（http://tigrblast.tigr.org/tgi）。

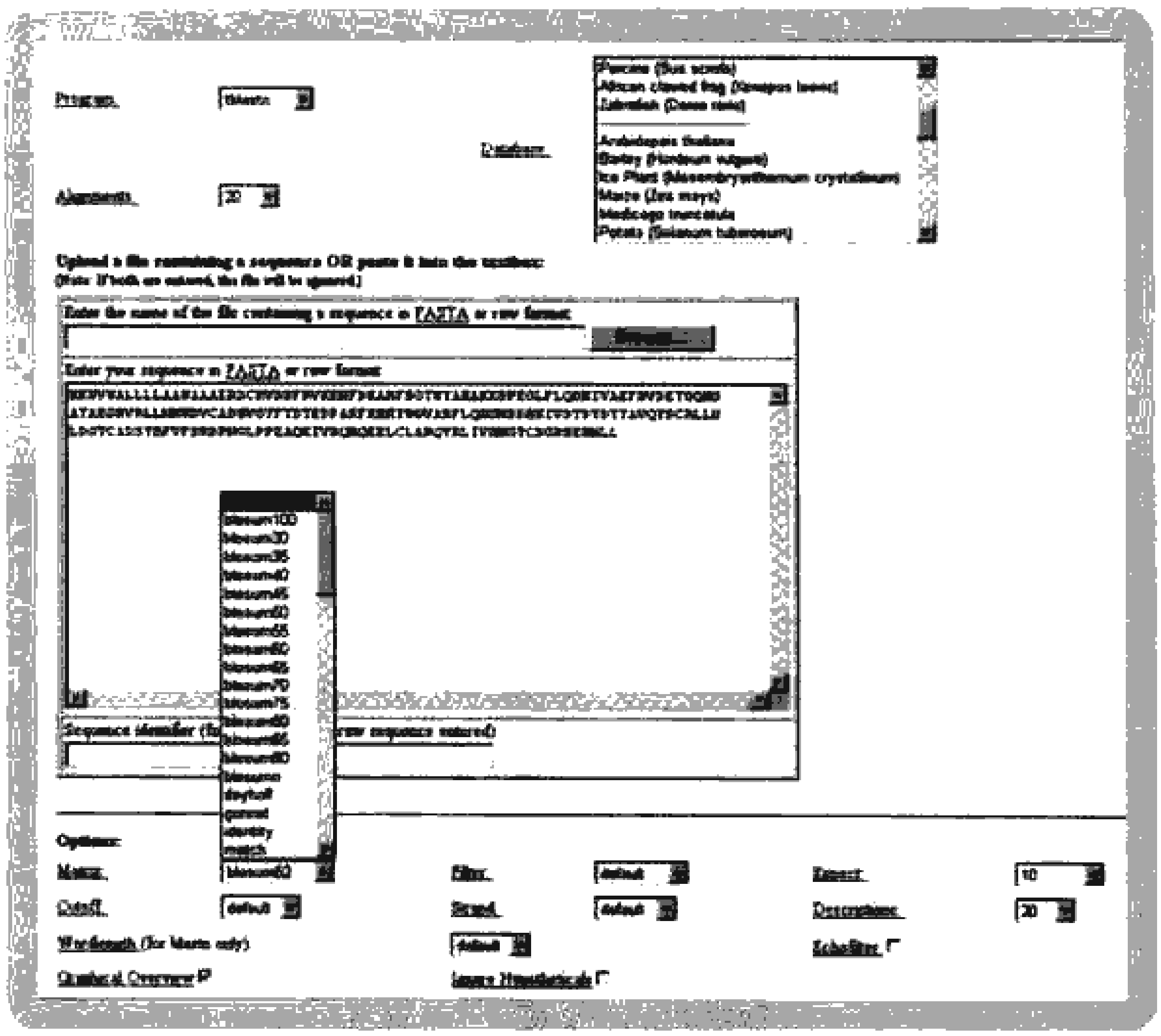


图 5-3 TIGR 上的 BLAST 搜索可以选择不同物种的数据库，就像从许多打分矩阵中选择一个合适的一样。注意这里的输入序列必须是 FASTA 格式的

Ensembl 和 TRIGR 比对搜索是很多其他数据库搜索的典型代表，它们都允许选择以不同的方式输入查询的条件，可以选择多种算法，而且提供一些包含其他有用信息的网站链接作为输出结果的一部分。

5.2.2 特定分子的比对搜索站点

一些比对搜索的网站专门提供有关某些特定生物分子信息的查询（表 5-2）。例如，NCBI 网页上就含有用来分析免疫球蛋白的 igBLAST。igBLAST 结果可以显示与查询

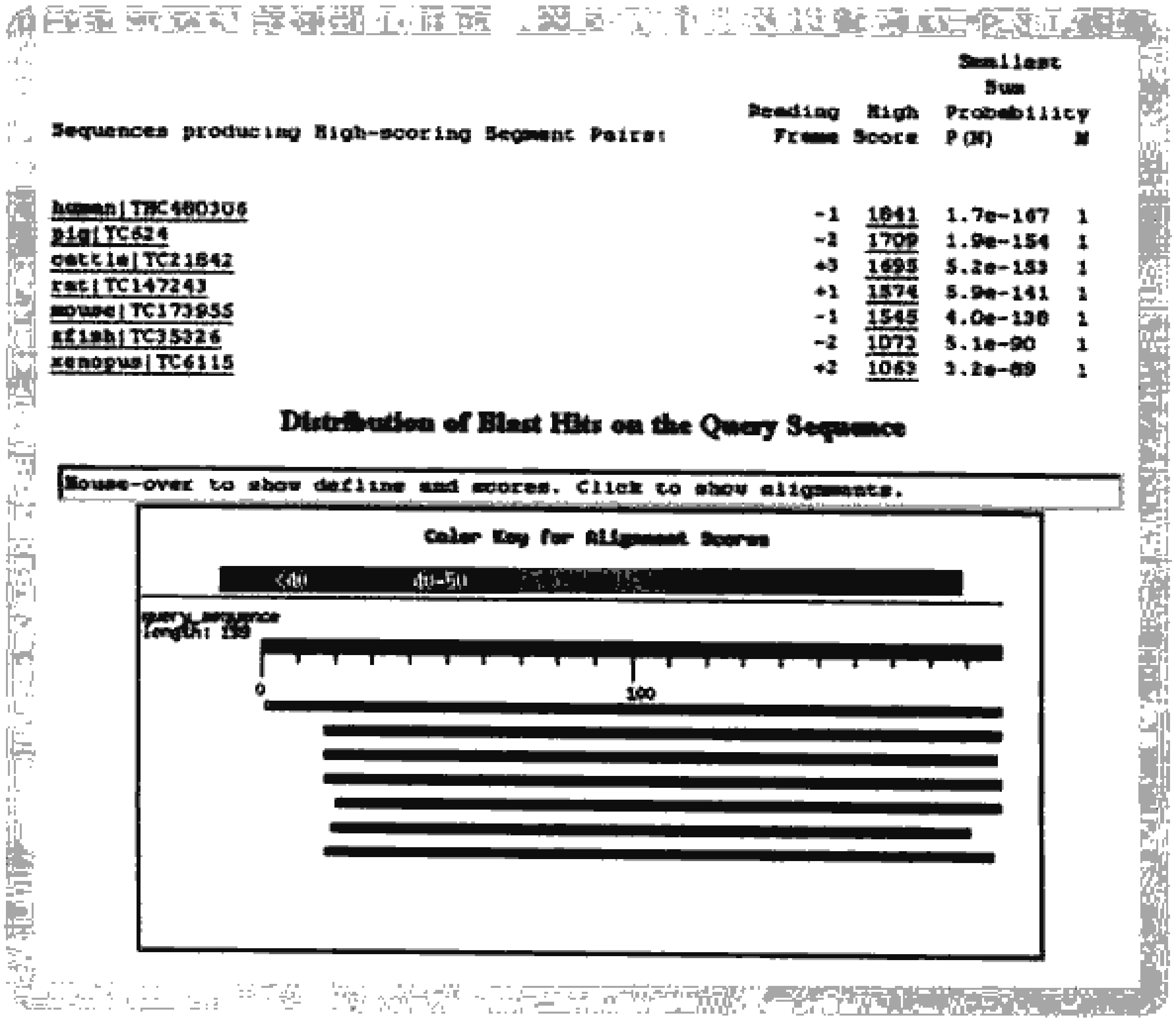


图 5-4 TIGR 上的 BLAST 结果和 NCBI 上的 BLAST 结果很相似。注意返回的匹配物种很多。一般一个物种只有一个匹配，因为冗余的序列在一个分类数据库中用同一个分类号

TC—潜在的一致序列；THC—潜在的人类一致序列

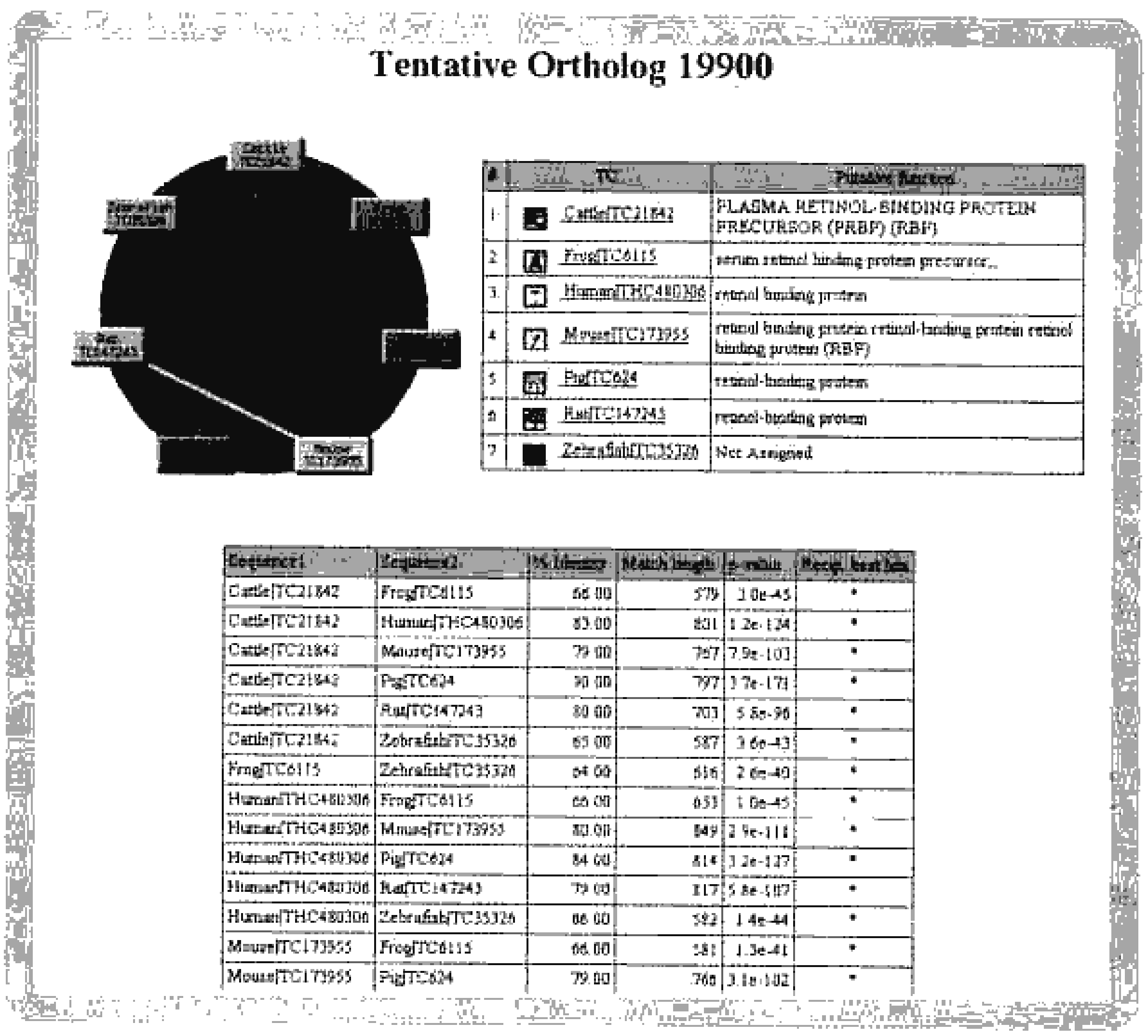


图 5-5 TIGR 中搜索的 BLAST 结果含有一个潜在同源组（TOGs）的链接，如 RBPs。这个 TOG 包含了 DNA 序列的多序列比对（未显示）

序列匹配性最好的3个V基因、2个D区域以及2个J基因。而且还能够给免疫球蛋白的功能域进行注释。最后，igBLAST匹配显示V区域最接近的比对结果。

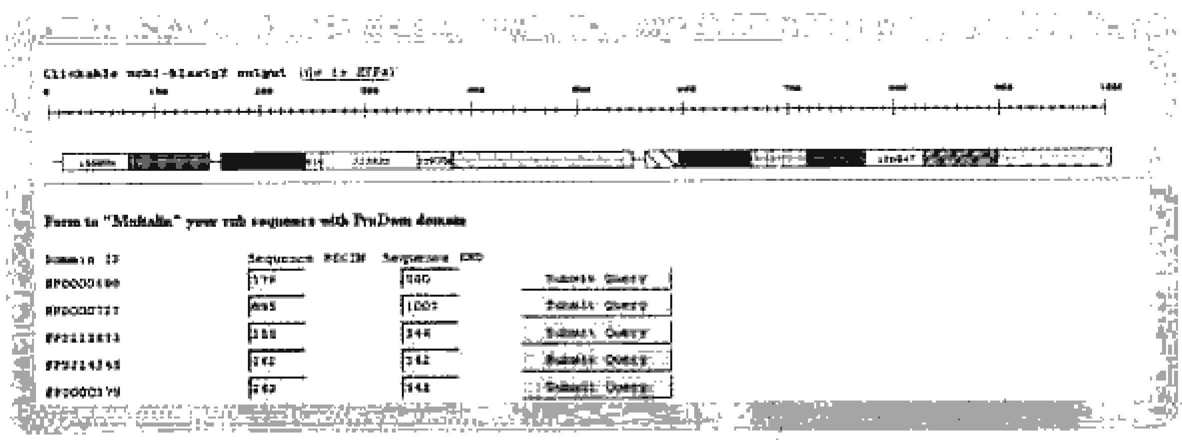
表 5-2 特定分子比对和相关比对的网站

分 子	描 述	网 址
免疫球蛋白	IgBLAST	http://www.ncbi.nlm.nih.gov/BLAST/
质谱比对	质谱比对(WU BLAST 2.0算法)	http://dove.embl-heidelberg.de/Blast2/msblast.html
蛋白质	网络蛋白质序列分析	http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_blast.html
	EMBnet-CH 网络比对服务	http://kr.expasy.org/cgi-bin/BLASTEMBnet-CH.pl
	ProDom 比对(含图形界面等输出)	http://prodes.toulouse.inra.fr/prodom/doc/blast_form.html
单核苷酸多态性	At NCBI	http://www.ncbi.nlm.nih.gov/SNP/snpblastByChr.html
转录组装程序	基于EST的基因查找(WU-BLAST 2.0算法)	http://sapiens.wustl.edu/~zkan/TAP/
载体(VecScreen)	对检测载体污染的DNA序列很有帮助	http://www.ncbi.nlm.nih.gov/BLAST/

注：这些网站提供针对特异分子的数据库搜索。

另外，有些服务器专门用来搜索蛋白质的相关信息（而不是核酸）。ProDom比对服务器支持标准的比对搜索输入形式而且以类似于NCBI的形式显示查询结果，但是它另外包含了一些图形界面用来形象地显示蛋白质的结构域（图5-6）。

图 5-6 ProDom服务器(http://prodes.toulouse.inra.fr/prodom/2002.1/html/form.php)的BLAST搜索结果提供了蛋白质功能域的图形界面。这里，是用HIV-1 pol (NP_057849)进行搜索的。有关HIV-1的每个功能域的信息还有一些链接



5.2.3 专门的比对搜索服务器以及比对相关的算法

> 你能够在http://blast.wustl.edu上学到更多关于WU BLAST 2.0的知识。

在前面的章节中一直都在介绍NCBI的标准比对算法，但是目前还存在很多其他的比对算法。表5-3列举了其中的一些算法，如WU比对2.0算法。WU比对2.0算法首先由华盛顿大学开发，它与传统的NCBI比对算法存在一定的相关性，因为这两种算法都由起初的NCBI比对算法发展而来，而且它们都不允许序列比拼中出现缺口。但是WU比对的速度更快而且敏感性更高，而且它也提供了很多的选择，例如可以在一些数据库搜索中进行完整的史密斯-沃特曼比对。WU比对2.0算法目前运行于很多服务器中，包括EMBL/EBI（图5-7）和TIGR（图5-3）。另外，还有一些其他的服务器提供一些对典型的比对算法进行微小改动而产生的专门的比对算法（如日本的DNA数据库，见图5-8）。FASTA算法目前也被广泛地使用（表5-3）（Pearson和Lipman，1988）。

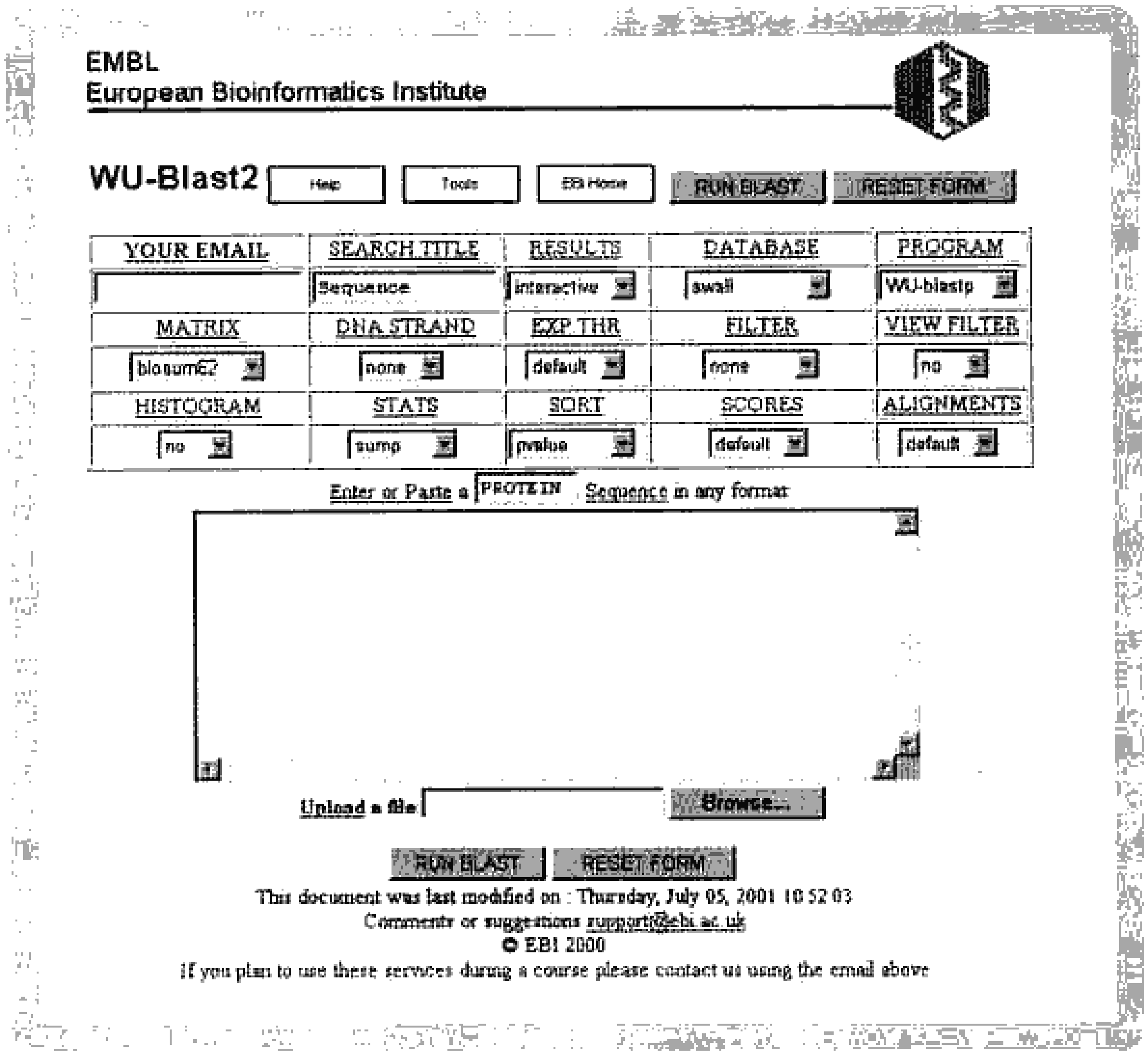


图 5-7 欧洲生物信息学研究所（EBI；http://www.ebi.ac.uk/）提供了一个 WU-BLAST2 站点（http://www2.ebi.ac.uk/blast2/）

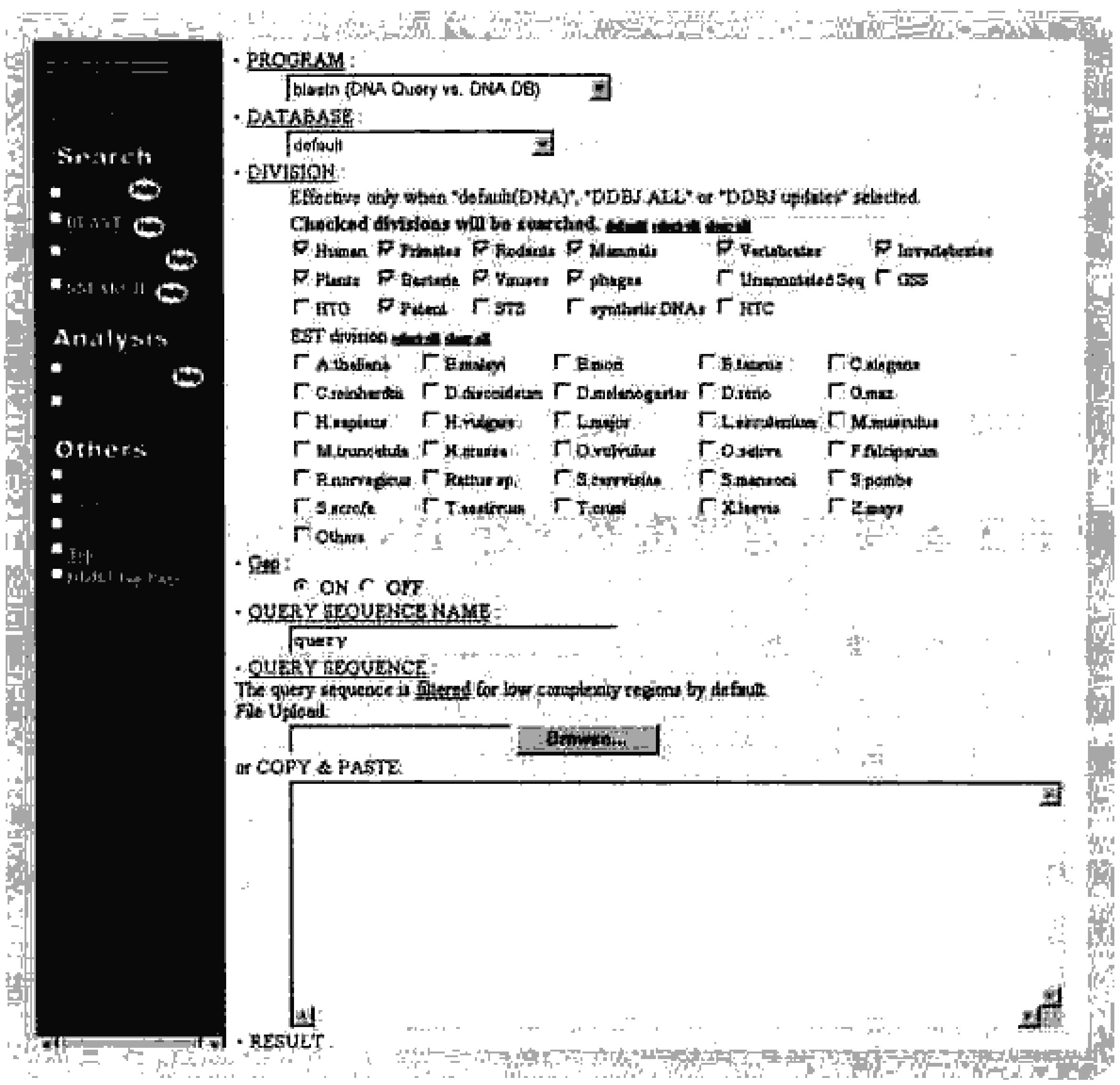


图 5-8 日本 DNA 数据库网站的 BLAST 页面（http://www.ddbj.nig.ac.jp/E-mail/homology.html）

表5-3 因特网上专门的比对算法和特异的比对相关算法

算法	描述	网址
PSI-BLAST	位点特异性反复比对	NCBI
PHI-BLAST	模块启始的比对	NCBI
MegaBLAST	用于大量DNA数据搜索	NCBI
RPS-BLAST	见第6章	NCBI
NCBI BLAST2	EMBL/欧洲生物信息研究中心	http://www2.ebi.ac.uk/blastall/
FASTA3	EMBL/欧洲生物信息研究中心	http://www2.ebi.ac.uk/fasta3/
WU BLAST 2.0	EMBL/欧洲生物信息研究中心	http://www2.ebi.ac.uk/blast2/
WU BLAST 2.0	EMBL	http://dove.embl-heidelberg.de/Blast2/
WU BLAST 2.0	华盛顿大学的基因测序中心	http://genome.wustl.edu/gsc/Blast/client.pl
WU BLAST 2.0	巴斯德研究所	http://bioweb.pasteur.fr/seqanal/interfaces/blastwu.html
WU BLAST 2.0	Mouse Genome Informatics at Jackson Labs	http://mouseblast.informatics.jax.org/prototype/
WU BLAST 2.0	转录组装程序	http://sapiens.wustl.edu/~zkan/TAP/
WU BLAST 2.0	孟德尔生物信息组织	http://www.mendel.ac.uk/blastaccess.html
WU BLAST 2.0	基因研究中心	http://tigrblast.tigr.org/tgi/
BLAST	瑞士生物信息中心和瑞士科学计算中心	http://www.ch.embnet.org/software/BottomBLAST.html?
FASTA	日本DNA数据库	DDBJ
BLAST	日本DNA数据库	DDBJ
SSEARCH	日本DNA数据库	DDBJ
PSI-BLAST	日本DNA数据库	DDBJ
FASTA	来自维吉尼亚大学(比尔·皮尔逊)	http://alpha10.bioch.virginia.edu/fasta/
SSEARCH	来自维吉尼亚大学(比尔·皮尔逊)	http://alpha10.bioch.virginia.edu/fasta/
BSU	比对搜索更新	http://athena.bioc.uvic.ca/genomes/index.html#BSU
MSPcrunch	用于比对结果的分析	http://www.embnet.se/new/mspup.html
Sim4	比对cDNA和基因组DNA	http://pbil.univ-lyon1.fr/sim4.html
Sim4	从宾夕法尼亚州的生物信息学组织可下载的软件	http://bio.cse.psu.edu/
BLAT	比对相似的工具,用于比对cDNA和基因组DNA	http://genome.ucsc.edu

注：NCBI指http://www.ncbi.nlm.nih.gov/BLAST/；DDBJ指日本DNA数据库的搜索工具。(http://www.ddbj.nig.ac.jp/E-mail/homology.html)。

其他的一些数据库搜索算法（如比对相似的比拼工具、PSI比对以及Pfam等）所运用的搜索策略基本上不同于比对算法。下文将详细地讨论这些算法。

5.3 运用比对相似的工具快速地搜索基因组DNA序列

随着基因组数据库中基因数据的快速积累（第12～第17章），现在越来越多地需要用对应于表达转录产物的蛋白质或DNA序列来搜索这些数据库。这是一个很专门的问题，包含以下四个方面。

> 我们将会在第6章（基因表达）、第12章（基因组）、第16章和第17章（真核生物基因组）讨论外显子和内含子。

① 基因组DNA序列包括外显子（对应于编码区的DNA片段）和内含子（对应于基因内部的非编码区）。理想情况下，一个比对工具应该可以找到基因组内基因的外显子。

② 基因组DNA序列往往含有一些测序引入的误差，这些误差引起的DNA序列错误应该加以考虑。

③ 在某些情况下，往往需要比较两个亲缘关系相近的物种的基因组 DNA 序列的相似性，如人和老鼠。

④ 这些算法应该能够找出 DNA 序列间的微小差异，如单核苷酸多态性（见第 18 章）。

现在已经发展了一些比对相似的算法可以不同程度地满足上述需要，列举如下。

① NCBI 比对网站包含一种大型比对算法，它类似于 blastn 算法并且被专门优化后用来快速对大量 DNA 序列数据进行比对搜索。

② Sim4 用一种比对相似的算法来搜索高得分区域的序列比拼（第 4 章）而且可以对这些区域进行双向延伸。一个动态的程序算法可以区分某一高得分率的比拼的序列片段是否代表某一基因。例如，这种程序可以用来搜索潜在的外显子和内含子连接处的剪接信号。

③ 用于序列搜索和比对的哈希算法（Hashing Algorithm）可以快速地搜索大型 DNA 数据库。它能够将一个含有数以亿计的碱基对的 DNA 数据库转换成一个哈希散列，这样就可以快速地搜索序列匹配了。

④ 比对相似的算法工具专门用于快速搜索基因组 DNA 数据库。正如用于序列搜索和比对的哈希算法一样，比对相似的算法工具在某些方面与比对工具很相像。比对首先将待比对的序列转化为字符串，然后用这些字符串去搜索数据库，并以一定的得分阈值来限制输出的结果。而比对相似的算法工具则将整个基因组 DNA 数据库中的序列全部转化为具有索引的字符串。这些字符串含有基因组内所有的非重叠的 11 个核苷酸片段（DNA 重复序列除外）。然后比对相似的算法工具从这些字符串中进行搜索比对以找到合适的匹配。

比对相似的算法工具还具有下列一些特点。

① 比对可以引发两个延伸搜索，而比对相似的算法工具可以引发多个延伸搜索。

② 比对相似的算法工具主要用于搜索序列相似性较高（95 或者以上）的匹配。虽然它在某种程度上与大型比对、Sim4，以及用于序列搜索和比对的哈希算法相似，但是它具有更高的搜索速度。

③ 比对相似的算法工具可以搜索内含子和外显子的边界序列，因此，可以用于建立基因结构的模型。它仅用从 mRNA 获得的核苷酸进行搜索（从生物学的观点来看，这是很恰当的），而不是仅仅搜索得分最高的匹配片段。

> 你能够在 http://pbil.univ-lyon1.fr/sim4.html 上得到 sim4，或者也可在 sim4 的作者（包括宾夕法尼亚大学的 Webb Miller）的网站（http://globin.cse.psu.edu）上下载得到这个程序。

> SSAHA 能够在 Ensemble 网站服务器（http://www.ensemble.org）上得到。SSAHA 的主页是 http://www.sanger.ac.uk/Software/analysis/SSAHA/。有一张哈希（Hash）表中包含了数据（如在一个 DNA 数据库中长度为 14 个核苷酸的序列的列表）和相关信息（如这些序列在 DNA 基因组中的位置）。

> BLAT 能够在 http://genome.ucsc.edu 上得到。这是主要的人类基因组浏览器之一，我们将会在第 17 章讨论它。

5.4 寻找远缘相关的蛋白质：位点特异性反复比对（PSI-BLAST）

很多同源的蛋白质都只含有有限的相似序列。这些蛋白质可能具有同一种三维结构（基于 X 射线衍射），但是如果进行序列比对，它们可能没有很明显的相似性。我们已经知道评分矩阵对不同进化关系的蛋白质匹配很敏感。例如，我们已经比较了 PAM250 和 PAM10 的评分矩阵（图 3-14 和图 3-15），并且看到了 PAM250 矩阵提供了一个更好的评分系统用

于探测远缘相关的蛋白质。因此，在进行序列比对时，可以通过调节序列矩阵来探测远缘相关的蛋白质。尽管如此，数据库中的很多蛋白质由于相关性太小，以致于很难用标准的 blastp 方法来检测。

你能够在 http://ncbi.nlm.nih.gov/BLAST 和其他一些服务器上得到 PSI-BLAST（见表5-5）。

位点特异性反复比对是一种专门化的比对，它往往比常规的比对算法更敏感。位点特异性反复比对主要用于搜索与我们感兴趣的蛋白远缘相关的蛋白。在很多情况下，当一个完整的基因组被测序后，往往需要寻找预测蛋白的同源物。这时，位点特异性反复比对就成为适当的选择。

位点特异性反复比对需要以下 5 个步骤。

① 首先用常规的 blastp 在目标数据库中进行比对搜索。本次搜索中主要用到一些常规的评分矩阵（如 BLOSUM62，该矩阵是默认的评分矩阵）。位点特异性反复比对也可以从一个蛋白质开始对 NCBI 网站上的数据库进行比对搜索。

我们已经在图4.12中的BLAST结果中见到了多序列比对，我们会在第10章讨论这个主题。

② 位点特异性反复比对从第①步得到的结果中构建多序列比对。然后为每个比对建立一个专门的搜索矩阵。

③ 用第②步中得到的定点评分矩阵再一次搜索原来的数据库（注意，这一次不要用第①步中的引物蛋白质）。

④ 位点特异性反复比对后用缺失比对的参数检验每个匹配的统计显著性。

⑤ 重复②～④步，一般要重复 5 次。在每次循环中都用上一次构建的多序列比对对数据库进行搜索。当新的结果不再出现或者程序明确指出不会再有新的结果出现时，可以停止比对循环。

PSI-BLAST 在概念上和逆向位点特异性 BLAST（RPS-BLAST）有很大的联系，RPS-BLAST 在第8章将会被详细讨论。PSI-BLAST 和 RPS-BLAST 都是用多序列比对和 PSSM，但是 RPS-BLAST 的目的是为了在一个预先设定的 PSSM 集中搜索一个蛋白质，然后找到保守蛋白质功能域。

下面举例说明位点特异性反复比对的过程和结果。进入 NCBI 网站，选择位点特异性反复比对（PSI-BLAST），然后输入人 RBP4 蛋白的序列号（NP_006735）。选择默认的参数，进行一次非冗余的比对搜索，这样就得到了 100 个匹配结果（这是 2001 年 12 月份的结果，所以这个结果肯定会有所变化）。其中有一半蛋白处于我们选择的阈值以上（默认的 $E=0.005$），通过进一步研究发现这些蛋白质都属于 RBP（来自不同的物种）。另外还有大约 50 个匹配低于我们默认的阈值：它们都没有很明显的 E 值。但是这些远缘的匹配中有一部分确实是 RBP，如它们与 RBP 具有相似的三维结构，并且具有作为载体的相似功能。其他蛋白则是一些病毒蛋白，很明显并非我们想要的结果。

你能够调整阈值，试着将 E 值改为 0.5 或 0.00005，看一下搜索结果有什么变化。如果将 E 值设得太低，就只能看到非常近源的物种。而如果将 E 值设得太高，就可能找到假阳性的同源匹配。

在这一步骤中，位点特异性反复比对与标准的 blastp 比对基本上没有什么区别，如都用到了一些氨基酸替换矩阵（如 BLOSUM45 等）。但是，位点特异性反复比对程序从第①步中的匹配中构建了多序列的比对。通过分析该多序列比对，位点特异性反复比对建立了定点评分矩阵，起初的蛋白质只是构建该多序列比对的一个起始点。

到目前为止，我们可以把比对结果的一部分作为多序列比对（图 5-9）。在图 5-9 的一列中，我们发现了氨基酸残基 R、I、K。总体来说，残基 R 和 K 经常会被替换掉，但是很少有 I 残基会被 R 或者 K 替换掉的情况发生。位点特异性反复比对的主要部分是它包

含有关多序列比对中特定氨基酸位点的突变频率的信息。这一信息就是构造比标准 BLOSUM 或者 PAM 更敏感的评分矩阵的基础。因而可以实现用更高的敏感度来搜索目标数据库。

```
730496    66   FTVDENGQMSATAKGRVRLFNNWDVCADMIGSFTDTEDPAKFKMKYWGVASFLQKGNDDH 125
200679    63   FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH 122
206589    34   FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH 93
2136812   2            MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 53
132408    65   FKIEDNGKTTATAKGRVRILDKLELCANMVGTFIETNDPAKYRMKYHGALAILERGLDDH 124
267584    44   FSVDESGKVTATAHGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAASYLQTGNDDH 103
267585    44   FSVDGSGKVTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH 103
8777608   63   FTIHEDGAMTATAKGRVIILNNWEMCADMMATFETTPDPAKFRMRYWGAASYLQTGNDDH 122
6687453   60   FKVEEDGTMTATAIGRVIILNNWEMCANMFGTFEDTEDPAKFKMKYWGAAAYLQTGYDDH 119
10697027  81   FKVQEDGTMTATATGRVIILNNWEMCANMFGTFEDTEEPARFKMKYWGAAAYLQTGYDDH 140
13645517  1                                MVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 32
13925316  38   FSVDGSGKMTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH 97
131649    65   YTVEEDGTMTASSKGRVKLFGFWVICADMAAQYTDPTTPAKMYMTYQGLASYLSSGGDNY 126
                                 ↑        ↑           ↑           ↑              ↑
                               R.I.K      C         D.E.T       K.R.T          N.L.Y.G
```

图 5-9 PSI-BLAST 搜索首先是一个标准的类似 blastp 的搜索。搜索后产生一个序列文件或 PSSM。PAM 或者 BLOSUM 矩阵描述了一个氨基酸被其他氨基酸替代的可能性，其依据是对上千个蛋白质的统计分析。PSSM 是专门针对每一个进行 PSI-BLAST 搜索的蛋白质产生的。这张图显示了最初的 PSI-BLAST 结果的一部分。箭头所指的位置代表了相应氨基酸匹配的例子和每个位点实际能够出现的氨基酸残基。有些位点是不变的（如上面的 C），有些位点则是这组蛋白质所特有的能够容忍的氨基酸替代残基（如 R、I、K）

对于一个序列长度为 L 的待查询蛋白质，位点特异性反复比对会构造一个 $L\times20$ 的评分矩阵。每个矩阵的所有行都具有长度 L。图 5-10 显示了我们用位点特异性反复比对搜索 RBP 4 蛋白中构造的评分矩阵的部分结果。每一列代表 20 种常见氨基酸中的一种。观察丙氨酸的得分（行 6、行 11、行 12、行 14～16，以及行 42）。在某些情况下丙氨酸得了 +4 分（这一分数与 BLOSUM62 的分矩阵中的分值一样，如图 3-17）。然而，丙氨酸也会在不同的部位偶尔得到 +2 分、+3 分、+5 分。这一点充分显示了基于待搜索蛋白而构建的得分矩阵的优越性。下面考虑位置 40 处的色氨酸，它是数百种脂蛋白共有的 GXW 结构中的一部分。这一位置处的色氨酸（以及其他色氨酸）得到了 +12 分，而位置 13 处的色氨酸只得了 +7 分。这些例子共同显示了位点特异性反复比对的一个主要的优势：定点得分矩阵能够更精确地反映不同位置处氨基酸发生置换的概率。

> PSI-BLAST 搜索的第一步和 blastp 搜索并不一样，PSI-BLAST 是根据基于组成的统计来计算查询蛋白和数据库序列的组成（Schaffer 等，2001）。打分矩阵可能不完全是 BLOSUM62 或者你选择的其他矩阵，但它们可能是非常相似的。

在位点特异性反复比对的第一个循环步骤中，我们用到了定点得分矩阵，并且得到了高于默认阈值（$E=0.005$）的匹配序列。我们一开始用到的蛋白质序列只是构建位点特异性反复中多序列比对的一个起点。冗余序列（含有至少 94% 相同序列）都被删除掉了。这样就保证了一组序列很相近的蛋白质不致造成我们构建的定点评分矩阵出现偏差。在上述的位点特异性反复比对搜索中，需要在序列内部插入空缺的序列比对会被忽略掉（Altschul 等，1997），这样就保证了定点评分矩阵的宽度与起始蛋白质序列长度一致。多序列比对矩阵有时候对一个查询能够有多个匹配，其他的则可能连一个都匹配不上。一个惟一的打分矩阵就是从多序列比对发展出来的。分数就是每个位点的 $\lg(Q_i/P_i)$，其中 Q_i 是残基 i 在该位点出

```
        A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
 1 M   -1 -2 -2 -3 -2 -1 -2 -3 -2  1  2 -2  6  0 -3 -2 -1 -2 -1  1
 2 K   -1  1  0  1 -4  2  4 -2  0 -3 -3  3 -2 -4 -1  0 -1 -3 -2 -3
 3 W   -3 -3 -4 -5 -3 -2 -3 -3 -3 -3 -2 -3 -2  1 -4 -3 -3 12  2 -3
 4 V    0 -3 -3 -4 -1 -3 -3 -4 -4  3  1 -3  1 -1 -3 -2  0 -3 -1  4
 5 W   -3 -3 -4 -5 -3 -2 -3 -3 -3 -3 -2 -3 -2  1 -4 -3 -3 12  2 -3
 6 A    5 -2 -2 -2 -1 -1 -1  0 -2 -2 -2 -1 -1 -3 -1  1  0 -3 -2  0
 7 L   -2 -2 -4 -4 -1 -2 -3 -4 -3  2  4 -3  2  0 -3 -3 -1 -2 -1  1
 8 L   -1 -3 -3 -4 -1 -3 -3 -4 -3  2  2 -3  1  3 -3 -2 -1 -2  0  3
 9 L   -1 -3 -4 -4 -1 -2 -3 -4 -3  2  4 -3  2  0 -3 -3 -1 -2 -1  2
10 L   -2 -2 -4 -4 -1 -2 -3 -4 -3  2  4 -3  2  0 -3 -3 -1 -2 -1  1
11 A    5 -2 -2 -2 -1 -1 -1  0 -2 -2 -2 -1 -1 -3 -1  1  0 -3 -2  0
12 A    5 -2 -2 -2 -1 -1 -1  0 -2 -2 -2 -1 -1 -3 -1  1  0 -3 -2  0
13 W   -2 -3 -4 -4 -2 -2 -3 -4 -3  1  4 -3  2  1 -3 -3 -2  7  0  0
14 A    3 -2 -1 -2 -1 -1 -2  4 -2 -2 -2 -1 -2 -3 -1  1 -1 -3 -3 -1
15 A    2 -1  0 -1 -2  2  0  2 -1 -3 -3  0 -2 -3 -1  3  0 -3 -2 -2
16 A    4 -2 -1 -2 -1 -1 -1  3 -2 -2 -2 -1 -1 -3 -1  1  0 -3 -2 -1
...
37 S    2 -1  0 -1 -1  0  0  0 -1 -2 -3  0 -2 -3 -1  4  1 -3 -2 -2
38 G    0 -3 -1 -2 -3 -2 -2  6 -2 -4 -4 -2 -3 -4 -2  0 -2 -3 -3 -4
39 T    0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -3 -2  0
40 W   -3 -3 -4 -5 -3 -2 -3 -3 -3 -3 -2 -3 -2  1 -4 -3 -3 12  2 -3
41 Y   -2 -2 -2 -3 -3 -2 -2 -3  2 -2 -1 -2 -1  3 -3 -2 -2  2  7 -1
42 A    4 -2 -2 -2 -1 -1 -1  0 -2 -2 -2 -1 -1 -3 -1  1  0 -3 -2  0
...
```

图 5-10 用 RBP4（NP _ 006735）进行 PSI-BLAST 搜索的 PSSM 的一部分。查询蛋白的 199 个残基在第一列上，20 个氨基酸则在第一排。注意一个给定的氨基酸（如丙氨酸）的打分是能够变化的（比较 A14、A15 和 A16，各自的分值为 3、2、4）。第 40 号残基的色氨酸在几百个 lipocalin 蛋白质中是不变的。将 W40、W3 或者 W5（每个的分值都是 12）与 W13 比较（+7）：在 W3、W5 和 W40 位点上匹配获得的分值更高，相应的不匹配的罚分则更高。像这样的 PSSM 使得 PSI-BLAST 搜索比标准的 blastp 搜索具有更高的敏感性

现的估计概率，P_i是这个残基的背景概率。

位点特异性反复比对确定的惟一打分矩阵接下来用于进行反复查找。按下按钮“进行位点特异性反复比对步骤 2”就开始用刚才的打分矩阵进行重复查找了，找到的新蛋白质就添加到比对中去。在第二个循环步骤中（图 5-11），数据库中匹配的数目上升到了 173，包括 48 个在阈值之上的新成员（表 5-4）。有时如果 PSSM（特定位点记分矩阵）改变了，蛋白质比对会落在阈值之下。

表 5-4 用人视网膜结合蛋白（NP _ 006735）进行 PSI-BLAST 的结果

循环	命中次数	高于阈值的命中次数	新增的	低于阈值的命中次数
1	104	49	—	—
2	173	96	48	1
3	236	178	85	3
4	301	240	64	2
5	344	283	43	0
6	342	298	17	2
7	378	310	12	0
8	382	320	11	1
9	391	330	11	1
10	372	336	8	2
11	370	337	2	1
12	371	338	1	0

注：第一步代表了最初的搜索结果，与标准的 blastp 搜索相当，接下来的 PSI-BLAST 循环找到更多的高于阈值（$E=0.005$）的匹配；当循环次数大于 12 时，基本上不会再找到更多的匹配结果。

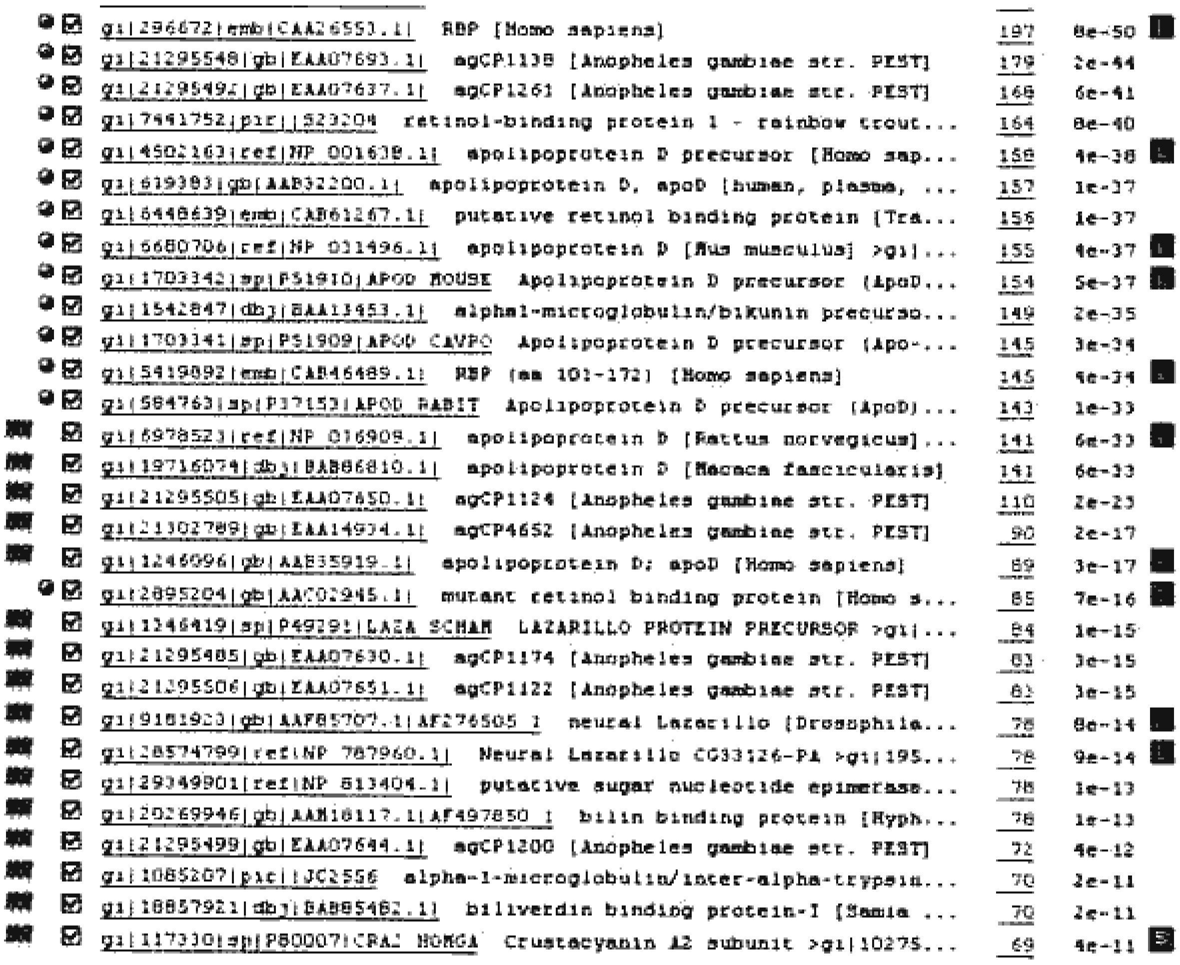


图 5-11 PSI-BLAST 搜索（步骤二）的结果的一部分。显示的是前面步骤中的匹配（左边球形标记）和新搜索到的序列。注意这里面包括一些称为 lipocalin 的蛋白质，表明它们可能是真正的 RBP 同源蛋白

继续这个反复过程直到程序出现“收敛了”这个消息框。这表明所有的数据库匹配都找完了，位点特异性反复比对完成了。

这个搜索查到了什么呢？在一系列位点特异性反复比对后，在数据库中找到了上百个其他的匹配，很多远缘的蛋白质也比对上了。我们能理解如何通过用老鼠载脂蛋白 D 和查询（RBP4）之间的比对来增加查找的敏感性。在第一轮位点特异性反复比对中，位分数为 46，期望值为 2×10^{-4}，150 个残基的比对中有 40 个是一致的和存在 37 个间隔。第二个和第三轮循环之后，位分数继续增长（先是到 140，然后到 159），E 值下降到 10^{-32} 再到 10^{-38}，比对的长度则先增长到 176 个残基，随后又轻微地下降到 170 个残基，而间隔数则下降得很厉害。在第二轮循环中发生了一个需要注意的变化，即比对的两个蛋白的氨基端都被包括进来了［图 5-12(a)，图 5-12(b)］。第三轮循环中有另外一个要注意的变化，即不断改变的打分矩阵关掉了一些大的间隔。

可以通过把数据库中的每一个脂质运载蛋白想像成为空间中的一个个点来理解位点特异性反复比对的过程（图 5-13）。最初用 RBP4 的查找发现了一些其他的 RBP 同源蛋白和载脂蛋白 D。位点特异性反复比对的 PSSM 可以发现和载脂蛋白 D 相关的一些其他脂质运载蛋白。老鼠的气味结合蛋白通过 RBP 的 blastp 搜索没有能够被发现，但是用位点特异性反复比对却可以。

另外一个例子可以用来说明位点特异性反复比对的有效性，即人类液泡蛋白——分类蛋白 45（h-vps45；NP_009189）（它是 Sec1 转运家族中的成员）。blastp 去冗余搜索后结果有 66 个匹配的 E 值在 0.002 以下。其他 E 值高于 0.002 的匹配似乎并不是这个蛋白质的相关同源蛋白。进行位点特异性反复比对搜索，一开始有 69 个匹配（与 blastp 去冗余搜索相比较），但是第一次循环有 87 个匹配，第二次循环有 94 个匹配，第三次循环后就收敛了（91 个匹配）。经过检查，有真正的相关蛋白质。

位点特异性反复比对搜索循环的次数和这个程序运行所调用的数据库中的匹配数目

(a)

```
Score = 46.2 bits (108), Expect = 2e-04
Identities = 40/150 (26%), Positives = 70/150 (46%), Gaps = 37/150 (24%)

Query: 27   VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC 86
             V+ENFD  ++ G WY + +K P      + I A +S+ E G +    K          ++
Sbjct: 33   VQENFDVKKYLGRWYEI-EKIPASFEKGNCIQANYSLMENGNIEVLNK---------ELS 82

Query: 87   ADMVGTF---------TDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCR 137
              D  GT          ++  +PAK +++++ +          +WI+ TDY+ YA+ YSC
Sbjct: 83   PD--GTMNQVKGEAKQSNVSEPAKLEVQFFPLMP-----PAPYWILATDYENYALVYSCT 135

Query: 138  ----LLNLDGTCADSYSFVFSRDPNGLPPE 163
                 L ++D      + ++  R+P  LPPE
Sbjct: 136  TFFWLFHVD------FFWILGRNPY-LPPE 158
```

(b)

```
Score =  140 bits (353), Expect = 1e-32
Identities = 45/176 (25%), Positives = 78/176 (43%), Gaps = 33/176 (18%)

Query: 4    VWALLLLAAWAAAERDCRVSSF--------RVKENFDKARFSGTWYAMAKKDPEGLFLQD 55
             V  L+ LA  A      +  +F         V+ENFD  ++ G WY + +K P      +
Sbjct: 2    VTMLMFLATLAGLFTTAKGQNFHLGKCPSPPVQENFDVKKYLGRWYEI-EKIPASFEKGN 60

Query: 56   NIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMV---GTFTDTEDPAKFKMKYWGVASF 112
              I A +S+ E G +    K       + D   + V      ++   +PAK +++++ +
Sbjct: 61   CIQANYSLMENGNIEVLNKEL-----SPDGTMNQVKGEAKQSNVSEPAKLEVQFFPL--- 112

Query: 113  LQKGNDDHWIVDTDYDTYAVQYSCR----LLNLDGTCADSYSFVFSRDPNGLPPEA 164
                    +WI+ TDY+ YA+ YSC     L ++D      + ++  R+P  LPPE
Sbjct: 113  --MPPAPYWILATDYENYALVYSCTTFFWLFHVD------FFWILGRNPY-LPPET 159
```

(c)

```
Score =  159 bits (404), Expect = 1e-38
Identities = 41/170 (24%), Positives = 69/170 (40%), Gaps = 19/170 (11%)

Query: 3    WVWALLLLAAWAAAERD--------CRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQ 54
              V  L+ LA  A             +  S   V+ENFD  ++ G WY + K
Sbjct: 1    MVTMLMFLATLAGLFTTAKGQNFHLGKCPSPPVQENFDVKKYLGRWYEIEKIPASFE-KG 59

Query: 55   DNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQ 114
             + I A +S+ E G +    K           V  +      ++   +PAK +++++ +
Sbjct: 60   NCIQANYSLMENGNIEVLNKELSPDGTMNQVKGE--AKQSNVSEPAKLEVQFFPL----- 112

Query: 115  KGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEA 164
                  +WI+ TDY+ YA+ YSC             + ++  R+P  LPPE
Sbjct: 113  MPPAPYWILATDYENYALVYSCTTFFWL--FHVDFFWILGRNPY-LPPET 159
```

图 5-12 PSI-BLAST搜索通过PSSM的渐进循环找到一些远源蛋白。(a)用RBP4(NP_006735)进行查询能够在第一步搜找到老鼠载脂蛋白D(NP_031496)。当搜索过程到了第二步(b)和第三步(c),分值越来越高,期望阈值降低,比对的空格数目也减少了

有关系。每经过一个位点特异性反复比对循环,返回的结果就显示哪些序列匹配输入的PSSM。

在一个相关的方法中,Schaffer等(2001)将PSI-BLAST搜索的假阳性数目对真匹配作图得到了一条"敏感性曲线"。他们利用这个图来估计不同参数下PSI-BLAST的准确率。

5.4.1 位点特异性反复比对的结果评估

有好几个途径来评估位点特异性反复比对的结果。当查询是在如SwissProt这样的一些大型数据库中进行的时候,特定位点记分矩阵(PSSM)能够在序列顺序打乱或者不同的数据库版本中进行搜索。这个时候,位点特异性反复比对的期望值就不是那么重要了(Altschul等,1997)。

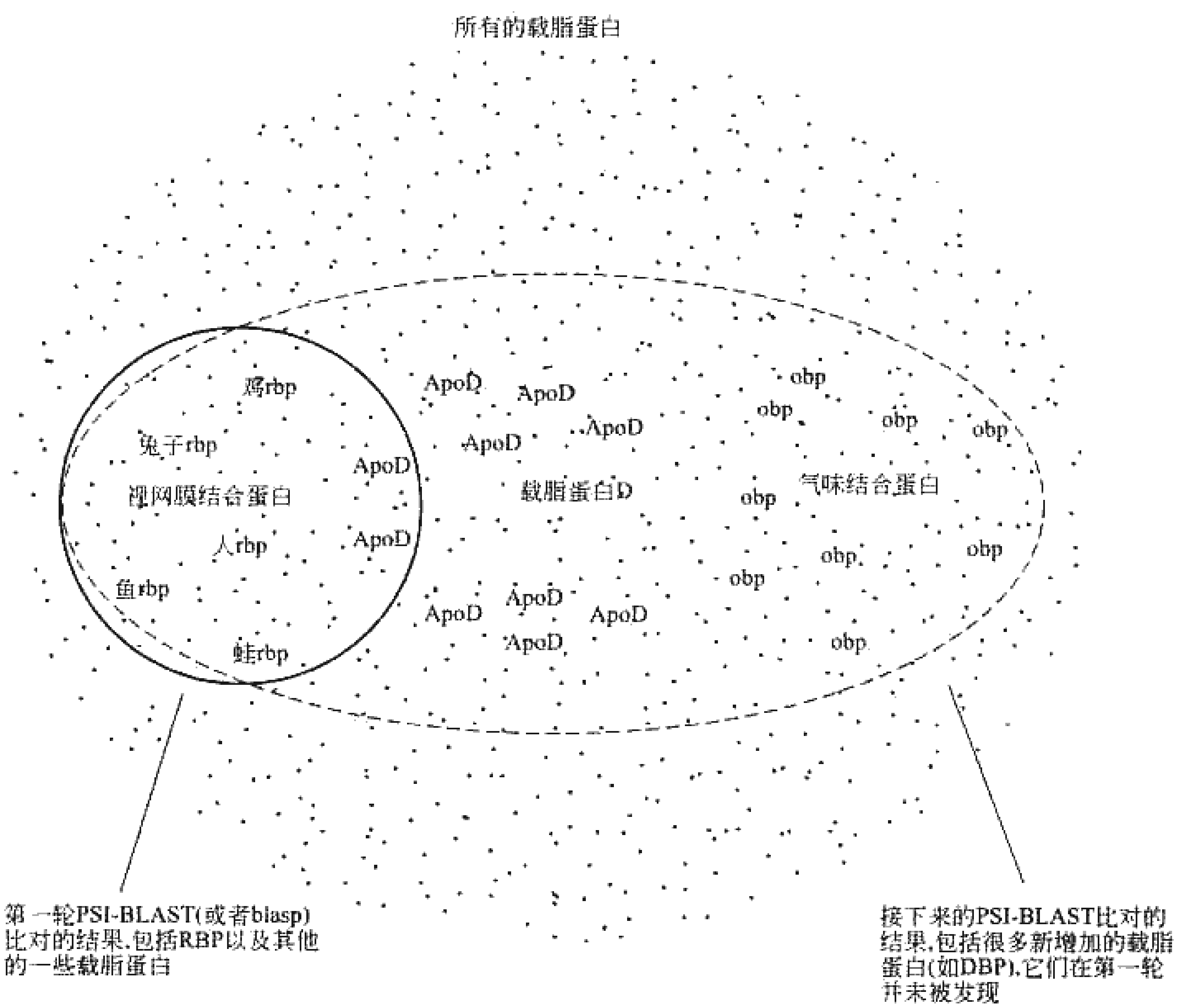


图 5-13 PSI-BLAST 算法能够找到具有相对较低序列一致性的同源匹配，从而增加数据库查询的敏感性。在这个图中，每一个点代表一个蛋白质，有些是标记了的 rbp（维生素 A 结合蛋白）、ApoD（载脂蛋白 D）或者 obp（气味结合蛋白）。由于它们都属于 lipocalin 家族，所有这些蛋白质都是同源的。用 rbp 进行标准的 blastp 搜索就会得到一些和 rbp 有序列一致性的近缘匹配，结果中还会有类似 ApoD 的 lipocalin 蛋白匹配。然而，许多其他的 lipocalin 蛋白（如 obp）就不能被找到。标准 BLAST 搜索敏感性差的最主要原因在于对标准 PAM 和 BLOSUM 打分矩阵的依赖。在一个 PSI-BLAST 搜索过程中，PSSM 针对输入的搜索序列（如 RBP）的匹配组产生一个打分系统。PSI-BLAST 搜索最初的结果是和标准 BLAST 搜索结果是一致的，PSI-BLAST 接下来的步骤（用椭圆形虚线表示）是用定制的打分矩阵来搜索更多的较远的同源蛋白

还有一种方法，有些研究小组将位点特异性反复比对搜索得到的结果和对同源蛋白（只有有限的一致蛋白）刚性结构分析得到的结果进行了比较。Park 和他的同事们（1998）用了蛋白质结构分类数据库（SCOP）。他们发现用 *E* 值为 0.005（*E* 值的默认值就是 0.005）的位点特异性反复比对搜索得到的结果更为准确。他们通过评估 432680 个可能的匹配中有多少个错误的预测来估计位点特异性反复比对匹配结果的假阳性率。在低假阳性率（1/100000）下，位点特异性反复比对在数据库中发现了 27% 的同源匹配；在较高的假阳性率（1/1000）下，位点特异性反复比对发现了 44% 的同源匹配。这个结果可以和 SAM-T98 相比拟，SAM-T98 是一个隐藏的马尔科夫模型

SCOP 数据库（第 9 章）的链接是 http://scop.mrc-lmb.cam.ac.uk/scop/。它是由 Cyrus chothia 和他的同事建立起来的。Park 等（1998）使用了 PDBD-J 数据库，这个数据库中的蛋白质氨基酸同源度≤40%。

Friedberg 等（2000）根据蛋白质结构比对（FSSP）和远源蛋白质序列比对（DAPS）（参见第9章）进行了折叠分类。他们研究了好几种 lipocalin 蛋白：牛的气味结合蛋白、小鼠的主要尿蛋白和后色胆素结合蛋白。

(Markov model) 程序（第10章）。位点特异性反复比对远比标准的缺位 BLAST 和 FASTA 算法敏感。

Friedberg 和他的同事们（2000）评估了位点特异性反复比对的准确性。他们从两个结构数据库中选了一些蛋白质。尽管这些用来比对的蛋白质的氨基酸的一致性低于25%，但用位点特异性反复比对成功地从123个已知结构蛋白质中搜索到了52个序列，而如果用 Smith-Waterman 算法则一对同源序列都没有搜索到。他们将位点特异性反复比对的结果和X射线晶体进行比对，发现平均大约44%的残基正确地比对了。

5.4.2 位点特异性反复比对的错误：破坏的问题

PSI-BLAST 在搜索蛋白质之间微弱但又具有生物学意义的联系的时候非常有用。导致 PSI-BLAST 搜索出现错误最主要的问题在于不断地找到一些无关的假阳性序列。当搜索的蛋白质（或者 PSI-BLAST 循环搜索之后产生的文件）中含有高度偏好性氨基酸组成的时候，这个问题可能会更加严重。一旦程序找到一个新序列，即使新序列的 E 值比设定的阈值稍高，也会被加入到下一个矩阵文件中，然后参与到下一次 PSI-BLAST 循环。如果这个匹配的蛋白质实际上不是最初查询序列的同源蛋白，那么 PSSM 就会被破坏。

我们将讹误（corruption）现象定义为在 PSI-BLAST 搜索5个循环之后，PSSM 至少产生了一个 E 值小于 10^{-4} 的假阳性匹配。这个定义是根据 Schaffer（2001）修改得到的。

有3个主要的方法可用来避免 PSI-BLAST 搜索查询的破坏问题。①能够运用一个过滤算法来将偏好性的氨基酸区域去除掉。这些“低熵”区域包括一些高度碱性的、酸性的或者富集脯氨酸的氨基酸序列。为了实现这个目的，NCBI 网站上提供了 SEG 程序，对在数据库中查询后得到的序列结果用过滤算法进行过滤。②能够将阈值从它的默认设置（如 $E=0.005$）调整为一个较低的值（如 $E=0.0001$）。这样就能够减少假阳性的出现，但是也会干扰正常匹配的发现。③能用肉眼来检查每一次 PSI-BLAST 循环。如图 5-11 所示，能够点击 PSI-BLAST 输出结果中每一个蛋白质的检验栏，然后将可疑序列除去。例如，查询的蛋白质可能含有一个无规卷曲区域，这就可能导致其他一些非同源蛋白质（如肌浆球蛋白）因为含有这个区域而获得较好的 E 值。

SEG 是由 Wootton 和 Federhen（1996）建立起来的，其地址为 ftp://ncbi.nlm.nih.gov/pub/seg/。

如果一个蛋白质有好几个模体（如同时含有激酶区域和 C2 区域），PSI-BLAST 搜索可以在数据库中找到和两个都有关的匹配。这个结果就需要谨慎对待了，不能随便得出激酶区域和 C2 区域相关的结论。在问题1中给出了这种情况的一个例子。在用 PSI-BLAST 对多功能域的蛋白质进行搜索的时候，用感兴趣的一个功能域来搜索会更有效，例如 HIV-1 pol 蛋白质的逆转录酶功能域。

5.5 模式识别 BLAST（PHI-BLAST）

很多时候感兴趣的蛋白质有特定的氨基酸残基模式或“信号”残基，能够用来帮助判断这个蛋白质属于哪一个家族。例如，信号残基可能是一个酶的活性位点，一个蛋白质家族的

结构或者功能域的氨基酸序列，或者甚至是未知功能的特征信号（如 lipocalin 中的 GXW）。模式识别 BLAST（PHI-BLAST；ϕ-BLAST）是一种特殊的 BLAST 程序，能够让查找得到既和查询匹配又和模式匹配的数据库结果（Zhang 等，1998）。PHI-BLAST 更加适用于和一种模式相关的短序列的简单数据库查询，因为这样的搜索结果可能会找到许多和查询的蛋白质无关的匹配或者一些任意的蛋白质。

PHI-BLAST 和 PSI-BLAST 都是在 NCBI 网页上发布的。它们是相互独立的程序，尽管 PHI-BLAST 搜索的输出结果能够作为 PSI-BLAST 的输入。PHI-BLAST 只需要执行一次（就像 blastp）而不需要多重反复。

> PHI-BLAST 的模式限制应用于原始数据库搜索（PSI-BLAST 的 PSSM 搜索不是这样）。

用人类蛋白质 RBP4 来对细菌的蛋白质序列进行 PHI-BLAST 查询搜索。结果（2002 年 8 月份的版本）是第一轮数据库中有 7 个匹配，所有匹配的细菌序列的分值都比默认阈值 E（0.005）差。细菌的 lipocalin 蛋白质和人类 RBP4 蛋白质相当远缘；有一个找出它们的办法就是用“bacterial lipocalin”作为关键字来搜索蛋白质。选择两个细菌 lipocalin 蛋白质和 RBP4 进行比对（图 5-14，第 10 章将讨论如何进行多序列比对）。这个比对会展示哪些氨基酸残基是 RBP4 和两个细菌蛋白质共有的。考虑到 GXW 模式几乎是所有的 lipocalin 蛋白质所共有的，我们能够定义一个氨基酸模式（或者信号）是 RBP4、这两个细菌的 lipocalin 蛋白质和其他细菌的 lipocalin 蛋白质所共有的。定义一个信号序列的目的是为了使得 PSI-BLAST 算法能够搜索到含有信号序列的蛋白质。

```
        1                                                  50
ecblc   MRLLPLVAAA TAAFLVVACS SPTPPRGVTV VNNFDAKRYL GTWYEIARFD
   vc   MRAIFLILCS V...LLNGCL G..MPESVKP VSDFELNNYL GKWYEVARLD
hsrbp   ~~~MKWVWAL LLLAAWAAAE RDCRVSSFRV KENFDKARFS GTWYAMAKKD
```

图 5-14 PHI-BLAST 搜索选择一个模式。将人类 RBP4（NP_006735）和两个细菌 lipocalin 蛋白一起进行多序列比对（大肠杆菌 P39281 和霍乱弧菌 Q08790）。把这三个序列放在一起进行多序列比对的目的是为了找到一个短而重要的氨基酸序列在这个蛋白质家族中普遍存在。用 GCG PileUp 程序进行了比对，图中显示了结果的一部分。通过观察发现典型的 lipocalin 蛋白具有固定不变的 GXW 模式（用深灰色表示）。这样就可以选择一个 PHI 模式，包括这些残基和一些其他的。例如，我们选择 GXW [YF] [EA] [IVLM]，GXW 之后的位点不是 Y 就是很近的 F，接下来就是 E 或 A，最后是疏水残基 I、V、M 和 L。用户能够通过反复试验来选择模式。接下来的数据库搜索的匹配必须包含有这个选择的模式

信号序列或者模式是怎么被定义的呢？我们并不希望所有的细菌 lipocalin 蛋白质的信号序列是完全一致的，因此会引进一些模糊的自由度。我们能够定义任何的模式，这可以图 5-14 的多序列比对为例子来说明，这里定义了模式 GXW [YF] [EA] [IVLM]。注意选择的模式一定不能出现得太频繁，这个算法只允许模式在数据库中发生的频率小于每 5000 个残基出现 1 次。一般来说，选择 4 个比较特殊的残基的模式或者平均背景频率≤5.8%的 3 个残基的模式都是可以接受的（Zhang 等，1998）。

> PHI-BLAST 模式的选择规则是根据 Prosite 字典得到的，具体见下面的链接：http://www.ncbi.nlm.nih.gov/blast/html/PHIsyntax.html。

BLAST 搜索的输出结果是一个含有这个特殊模式的蛋白质的数据库子集。经过对选择模式的研究后发现，每一个数据库中的匹配必须有一个 G，X 允许接着连上任何残基，W 表示这个模式的第三个氨基

酸必须是 W。接下来，[YF] 表示随后的氨基酸必须是 Y 或 F 中之一，选这个模式是因为用苯丙氨酸来代替酪氨酸相当常见（见第 3 章）。在接下来的位置上选择 [EA]，因为这两个氨基酸在图 5-14 的比对中出现了。最后，选择 [IVLM] 来和比对中的 3 个残基（I，V，M）再加上另外一个疏水残基（L）相对应，加上疏水残基 L 是因为有种直觉认为这个位置上会有疏水残基。

接下来登录 PSI-BLAST 网页（http://www.ncbi.nlm.nih.gov/BLAST）以使用 PSI-BLAST。在查询选项下面注意一个称为“PHI 模式”的框，在其中输入模式 GXW [YF] [EA] [IVLM]（图 5-15）。这个搜索的结果数据库中有许多匹配而不是只有先前的 5 个匹配（打分没有一个比阈值好），包括 6 个细菌 lipocalins 蛋白质（打分都比阈值好）。在接下来的 PSI-BLAST 过程中，不使用 PHI 模式而使用特定的 PSSM，就能成功地找到一大家族的细菌 lipocalin 蛋白质。

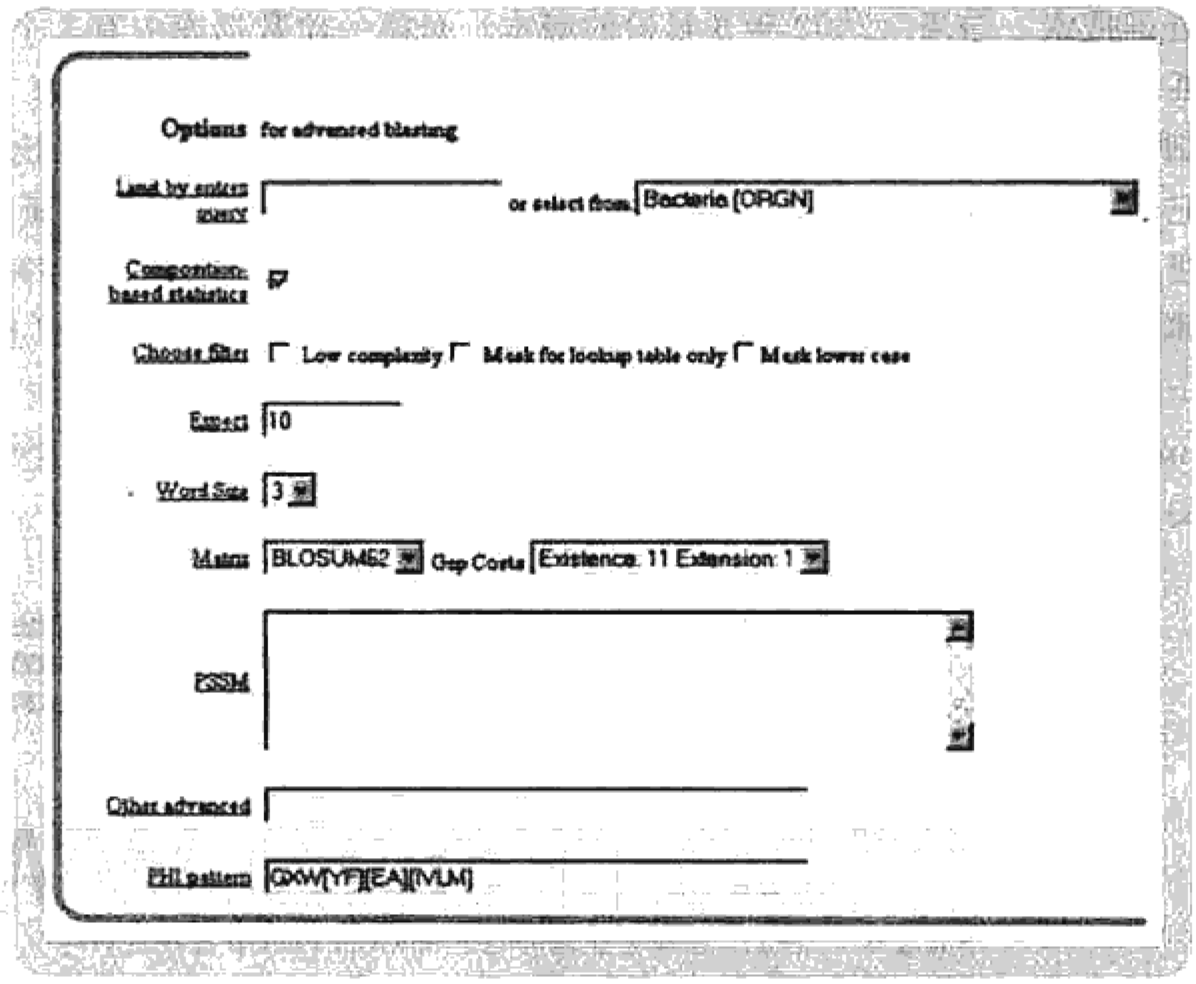


图 5-15 PHI-BLAST 搜索用几个氨基酸序列作为信号序列来进行数据库查询。PHI 模式的构建方法和 PROSITE（第 8 章）使用的构建方法是一样的。如果用户把这个模式输入选项中就能够进行 PHI-BLAST 搜索。增加了 PHI 模式常能够增加数据库中匹配的数目。通过选择一个蛋白质家族中的保守区域作为一个合适的模式输入就可以进行 PHI-BLAST 搜索了

PHI-BLAST 搜索的输出结果和 PSI-BLAST 的输出结果是一致的，除了显示的查询和每个数据库中和 PHI 模式匹配的序列信息外（图 5-16 所示的部分）。PHI-BLAST 算法根

```
>gi|2497702|sp|Q46036|BLC_CITFR   OUTER MEMBRANE LIPOPROTEIN BLC PRECURSOR
 gi|2121019|pir||I40710   outer membrane lipoprotein - Citrobacter freundii
 gi|717136|gb|AAC46456.1|   lipocalin precursor [Citrobacter freundii]
          Length = 177

 Score = 17.3 bits (52), Expect = 0.002
 Identities = 23/82 (28%), Positives = 36/82 (43%), Gaps = 1/82 (1%)

Query:   27  VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC 86
pattern 38              ******
             V  NFD  R+ GTWY +A+ D        D + A +S+ + G ++    KG         W
Sbjct:   30  VVNNFDAKRYLGTWYEIARFDHRFERGLDKVTATYSLRDDGGINVINKGYNPDREMWQK- 88

Query:   87  ADMVGTFTDTEDPAKFKMKYWG 108
              +      FT       A  K+ ++G
Sbjct:   89  TEGKAYFTGDPSTAALKVSFFG 110
```

图 5-16 PHI-BLAST 搜索的结果和标准 BLAST 或 PSI-BLAST 搜索的结果很相似，只不过多了选择模式的信息。这里，查询蛋白（RBP4，NP_006735）和数据库中的匹配（一个细菌 lipocalin 蛋白，Q46036）根据选择的模式（GXW [YF] [EA] [IVLM]）进行比对

据输入模式 A_0 及其上下游的 A_1 区域和 A_2 区域进行统计分析，并且利用序列间隔进行打分。分值 S_0、S_1 和 S_2 分别表示这些区域计算后的打分，PHI-BLAST 的分值根据 $S_0'=S_1+S_2$（忽略 S_0）来排序。这个比对统计和用来 blastp 搜索的统计有很大的关系（Zhang 等，1998）。

5.6 用 BLAST 来发现新基因

生物学中一个普遍的问题就是寻找新基因。传统上，基因和蛋白质的发现需要用分子生物学和生物化学技术。互补 DNA 可以用克隆的方法从基因文库中克隆得到，而蛋白质纯化后可以用酶活性等生物化学指标测序（虽然这种实验生物学方法仍然是最基础的）。生物信息学方法也能够为新基因的存在提供一些证据。从我们的目的来看，一个新基因就是指在数据库中发现的一些还没有被注释的 DNA 序列。你会为了下面这些理由去寻找新基因。

① 可能在一个感兴趣的特殊的物种中（如植物或者古细菌）想研究一个以前从未发现过的 lipocalin。

② 对 lipocalin 感兴趣，并且发现在大颊鼠的眼泪中有一个 lipocalin 蛋白质，那么是否存在一个基因编码人类眼泪中的 lipocalin 蛋白质呢？从目前来看，确实发现了一个。

③ 想知道是否细菌也有 lipocalin 蛋白质。如果这样，就可以深入了解这个转运蛋白质家族的进化了。

④ 在研究有关糖代谢的疾病时，作为研究的一部分，要研究在人类细胞系中的糖转运。你知道通过生物化学方法已经把葡萄糖转运器给纯化出来了，也知道在 GenBank 中存储有一个葡萄糖转运器基因或者蛋白质家族。当你克隆出所有已知的转运蛋白，将它们在细胞中表达后发现没有一个重组蛋白质能够转运糖分时，就会设想一定至少有一个转运蛋白没有发现。有无办法可通过搜索数据库来找到编码这个新转运蛋白质的人类基因吗？

⑤ 正在研究 HIV pol 蛋白质，尤其是它的逆转录酶区域。你想在真核生物蛋白质中找到这个区域的例子，但不想研究一个已知有这个模体的真核生物蛋白质，而更愿意去研究一个从没有描述过的新蛋白质。

解决这些问题最为常用的方法如图 5-17 所示。经过练习，这整个过程能够在几分钟内完成。首先在数据库搜索中输入一个已知蛋白质序列（图 5-17，步骤 1）。搜索最好在一个有 EST、基因组序列或者其他 DNA 序列的数据库中进行。一般在酵母（啤酒酵母）、大肠杆菌或者其他那些 DNA 或者蛋白质非冗余的数据库中搜索的结果都不是特别有效，这是因为上面这些数据库中的序列都较好地被注释了，所以也就很难再找到新基因了。一个可行的方法就是在表 5-1 中列出的那些 DNA 数据库中用 tblastn 搜索感兴趣的蛋白质。

接下来要做的就是小心地检查 tblastn 结果（图 5-17，步骤 2）。实践表明，很可能出现以下 3 种结果。

① 在一些数据库中得到的结果完全匹配或者几乎完全匹配。如果搜索的是人类 RBP4 蛋白质而数据库不是人类的，那么这个匹配就可能是已经发现的 RBP4 的同源蛋白。这就不是新发现了。

② 一些数据库的结果也非常匹配，而该数据库中编码这些蛋白质的 DNA 还没有被注释。这种情况下的序列就有可能是新基因。

③ 一些搜索结果并不是非常匹配。这就需要依靠经验来判读哪些数据库中的匹配是真正的匹配，哪些不是。真正的同源物应该和查询的蛋白质大小相似，并且它们也应该含有相同的模式，例如 GXW 就是 lipocalin 蛋白质典型的信号序列。

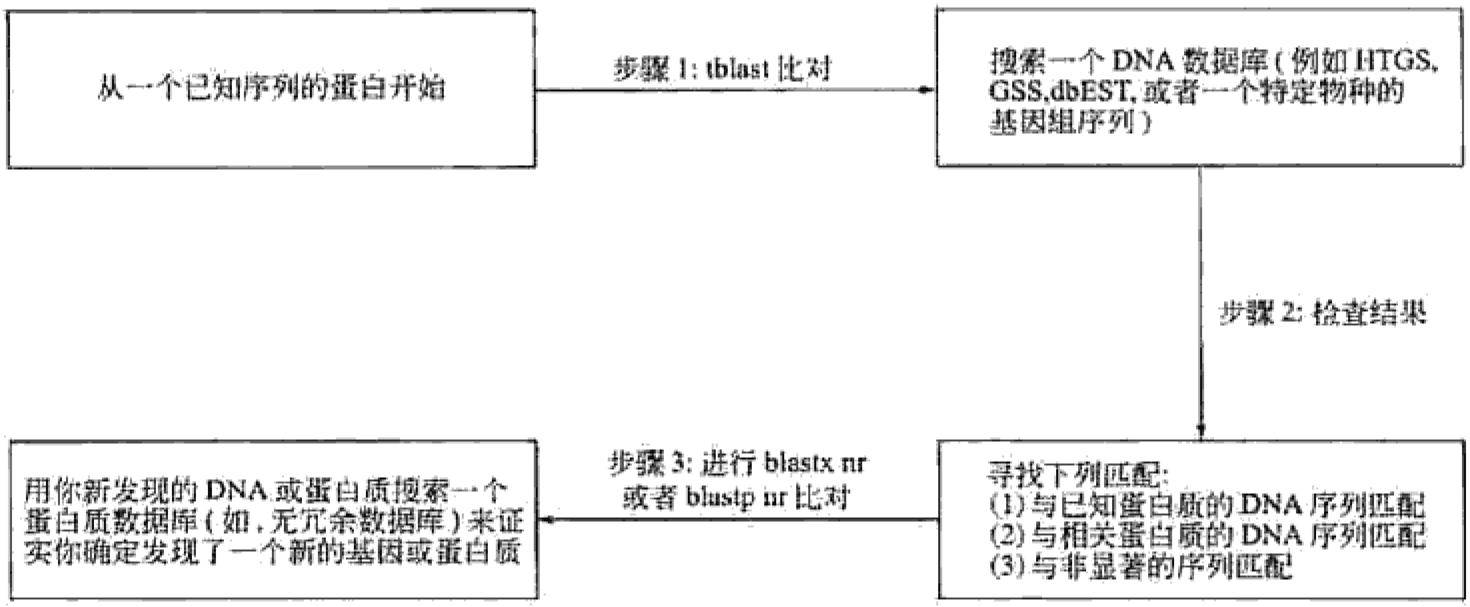


图 5-17 怎样通过 BLAST 搜索来发现新基因。首先从类似 RBP 的已知蛋白序列出发，一般在一个研究得较为透彻的物种（如人类、酵母或者大肠杆菌）的基因组中很难找到许多新基因。所以在几乎没有研究或者没有完全注释的物种数据库中搜索更为有效。Tblastn 搜索可能会有两种匹配结果：①与你查询蛋白所匹配的蛋白质已经被注释过了；②同源蛋白还没有被注释（新基因和其对应的新蛋白质）；③假设可能的新基因的 DNA 序列可以用 blastx 算法在非冗余数据库中进行搜索。从而可以确认这个 DNA 确实编码了一个蛋白质，这个蛋白质和以前注释过的蛋白质并不完全匹配

为了找到一个新 lipocalin 蛋白质，在一个没有完全测序的真核生物基因组中搜索一个大肠杆菌 lipocalin 蛋白（NP_290784）（http://www.ncbi.nlm.nih.gov/BLAST）。将数据库设置为“所有”，可显示几十个真核生物基因组（图 5-18）（或者也可以选择其他一些数据

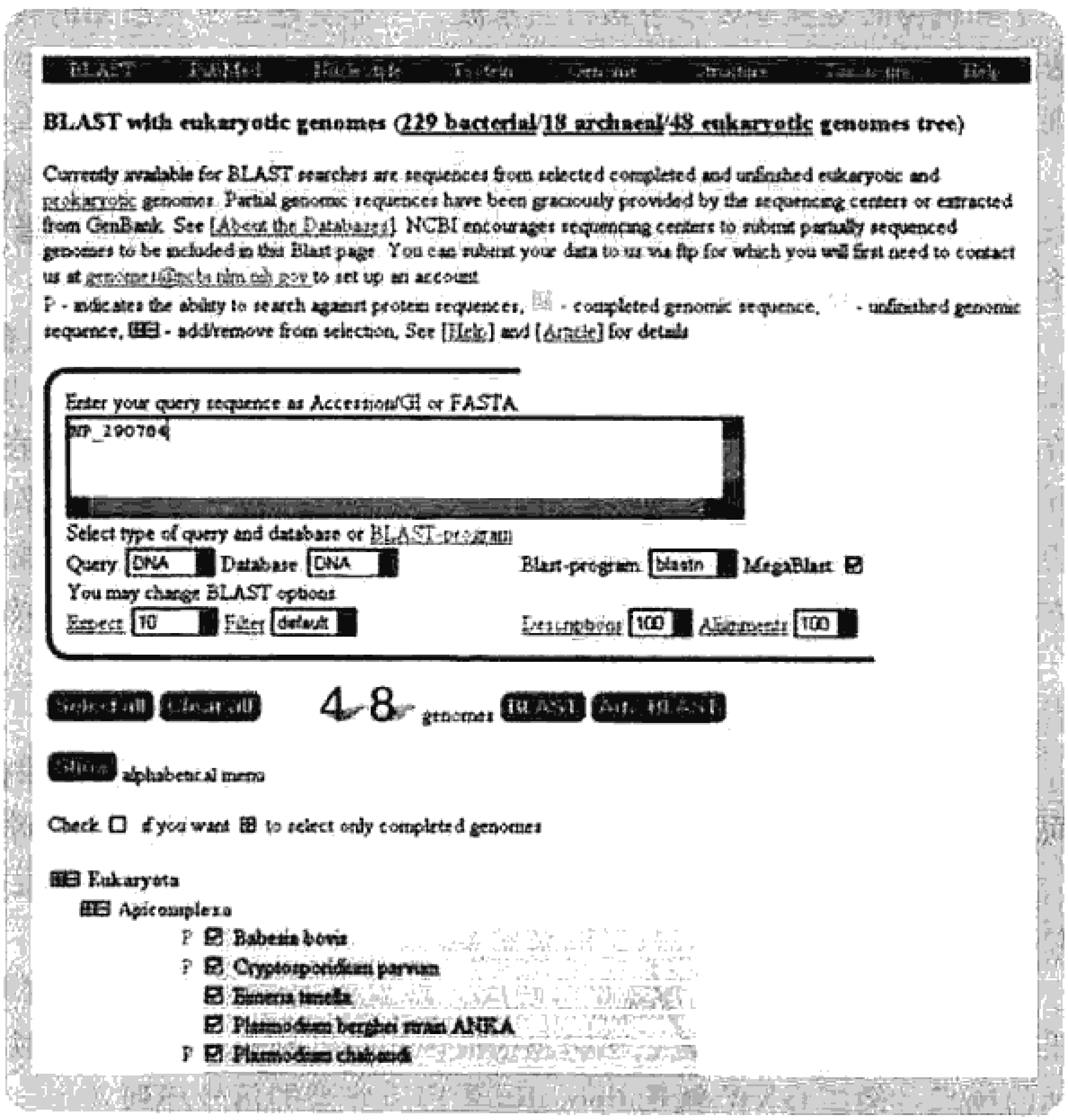


图 5-18 通过 BLAST 搜索寻找新基因的程序。这里用了“其他真核生物”基因组 BLAST 站点（http://www.ncbi.nlm.nih.gov/BLAST），输入人类 RBP4 蛋白质识别号，将程序设为 tblastn，选择物种选项。点击“选择所有”就可以得到一些物种的基因组了，如 *Aspergillus fumigatus*、*Entamoeba histolytica*、*Trypanosoma brucei* 和 *Plasmodium falciparum*。这个搜索的目的是为了找到与 RBP4 相关的蛋白质。据预测，在一些没有完成测序的基因组中有一些没有被发现的基因可编码 RBP4 蛋白

库进行搜索)。搜索的结果就是一组和细菌查询序列非常匹配的预测蛋白质(图 5-19),其中之一就是黏液菌(*Dictyostelium discoideum*)蛋白质。这可能是一个新蛋白质,因为之前在 *Dictyostelium* 基因组中没有人找出这个基因(或者其相应的蛋白质)。能够用数据库搜索来判断它是否真的是新蛋白质(如图 5-17,步骤 3)。用 Dictyostelium 蛋白质进行 blastp 搜索的结果表明它和其他一些细菌 lipocalin 蛋白质相当近源,但是它在 GenBank 数据库中还没有被注释过(如非常明确的描述)(图 5-20)。

(a)

```
                                                                    Score     E
Sequences producing significant alignments:                         (bits)  Value

gnl|ti 7165|44877221  19866907818432                                  176   5e-43
gnl|DictyConsortium 44689|jena JC1c122d04.s1  Clone JC1c122d...       149   4e-35
gnl|DictyConsortium 44689|jena JC1c276c01.r1  Clone JC1c276c...       116   4e-25
gnl|DictyConsortium 44689|jena JC1a241a10.s2  Clone JC1a241a...        99   7e-20
gnl|DictyConsortium 44689|jena JC1c236h06.r1  Clone JC1c236h...        94   3e-18
gnl|DictyConsortium 44689|iibsp.fa IIBCP1D0380  Dictyosteliu...        82   1e-14
gnl|ti 7165|47325132  19866908692623                                   68   2e-10
gnl|ti 7165|57669632  19866913818355                                   68   2e-10
gnl|ti 7165|55759626  19866909741201                                   63   5e-09
gnl|ti 7165|55559725  19866910183296                                   63   7e-09
gnl|DictyConsortium 44689|jena JC1b180c09.s1  Clone JC1b180c...        63   7e-09
gnl|ti 7165|60399996  19866914851608                                   61   3e-08
gnl|DictyConsortium 44689|jena JC1c45b10.r1  Clone JC1c45b10...        60   3e-08
gnl|DictyConsortium 44689|jena JC1b154f03.r1  Clone JC1b154f...        59   1e-07
```

(b)

```
>gnl|DictyConsortium 44689|jena JC1c276c01.r1 Clone JC1c276c01, reverse read,
          2001-01-30
         Length = 474

 Score =  116 bits (291), Expect = 4e-25
 Identities = 58/116 (50%), Positives = 75/116 (64%)
 Frame = +2

Query: 44  YEIARFDHRFERGLEKVTATYSLRDDGGLNVINKGYNPDRGHWQQSEGKAYFTGAPTRAA 103
            YEIAR +  +ER +  V+A Y+L  DG ++VIN GYN         +   GKA F      A
Sbjct: 116 YEIARLNFYYERDHNNVSAEYTLNKDGTISVINSGYNYVEKKRESLNGKALFVHGSHEAH 295

Query: 104 LKVSFFGPFYGGYNVIALDREYRHALVCGPDRDYLWILSRTPTISDEVKQEHLAVA 159
            LKVSFFGPFY GYNVIA+D +Y++ALV G     YLWILSR  +I D +K++ L +A
Sbjct: 296 LKVSFFGPFYAGYNVIAIDPDYKYALVAGRSLHYLWILSRETSIPDHIKKQYLQLA 463
```

图 5-19 (a) 在未完成测序的基因组中进行 tblastn 搜索的结果。(b) 算法。这个 *D. discoideum* 的蛋白质被预测是由一个基因组 DNA 片段来编码的。拷贝黏液菌蛋白质序列,为了判断它是否是一个"新"蛋白,我们将其作为输入项在一个非冗余数据库中进行 blastp 搜索

```
□>gi|23137293|ref|ZP_00119001.1|   hypothetical protein [Cytophaga hutchinsonii]
         Length = 179

 Score =  127 bits (319), Expect = 3e-29
 Identities = 66/116 (56%), Positives = 86/116 (74%)

Query: 1    YEIARLNFYYERDHNNVSAEYTLNKDGTISVINSGYNYVEKKRESLNGKALFVHGSHEAH 60
            YEIARL++ +E+++NNV+A Y+L +DG I V N GYN  ++K E   GKA  V    E
Sbjct: 46   YEIARLDYSWEKNLNNVTATYSLREDGKIKVDNKGYNVKKEKWEESVGKAAPVADPAEGR 105

Query: 61   LKVSFFGPFYAGYNVIAIDPDYKYALVAGRSLHYLWILSRETSIPDHIKKQYLQLA 116
            LKVSFFGPFYAGYNV+AID  Y YALVAG +  YLWILSR+ +IP+++K+ YL+ A
Sbjct: 106  LKVSFFGPFYAGYNVVAIDDAYTYALVAGENTKYLWILSRKKTIPENVKEAYLKKA 161
```

图 5-20 根据 blastp 搜索的结果,Dictyostelium 是一个新的 lipocalin 蛋白。数据库中和 lipocalin 最好的匹配来自其他生物(如 *Cytophaga hutchinsonii*)。这样,这种搜索方法就能够成功地在一个物种中找出一个没有注释也没有在 GenBank 数据库中储存的 lipocalin 蛋白

图 5-15 描述了一个相似的从任意物种中寻找基因的方法（然而，因为像人类、啤酒酵母和其他模式生物这样的基因组已经被很深入地研究过了，在这些物种中发现新基因变得越来越难了）。试着用 RBP4（用 FASTA 格式）和 tblastp 算法（见图 5-3）在 TIGR 数据库中进行搜索。如果这个方法没有结果，试着在不同的数据库中搜索或更换另外一个感兴趣的蛋白质进行搜索。

5.7 展望

随着 BLAST 搜索已经成为一个研究蛋白质和基因的基础工具（第 4 章），产生了许多特殊的 BLAST 应用。包括不同的算法（如 PSI-BLAST 的 PSSM）和特殊数据库（如许多生物的专门数据库）。PSI-BLAST 被广泛地用来在全基因组中寻找编码蛋白质（第 14～第 17 章）。PSI-BLAST 只能在 blastp 搜索能产生统计上有用的结果的情况下成功使用。

DNA 序列数据的指数级上升带来了海量的基因和蛋白质信息。BLAST 是在这些数据库中进行搜索的最基本的工具。对于回答有关在生命进化树上的基因家族方面的问题，BLAST 搜索的结果比查找文献的结果更为明确。本章描述了一些用 BLAST 对数据库进行搜索的方法。这些工具在今后多年内仍然对生物学相当重要，尤其是在基因组测序步伐继续加大的时候。

5.8 常见问题

对于任何生物信息学问题，必须要明确对数据库进行查询的目的，即要尽量去实现什么。一旦决定了目标，就能够选择合适的数据库和搜索算法。

对于 PSI-BLAST，最大的问题就是假阳性问题。一旦找到一个比期望阈值好的假序列，它就被包括在下一个循环的 PSSM 中了。这个循环又会找到假序列，并且增大数据库中匹配的数目。为了避免这个问题可采取以下措施。

① 在结果中检查是否有明显的假匹配。如果发现了假匹配，就将它们移除。

② 将期望值调整为一个合适的数值。

③ 为了判断 PSI-BLAST 的结果是否为假阳性，将这个序列作为输入序列在数据库中进行 BLAST“逆”搜索。

④ 进一步通过像在第 3 章描述的那样进行成对序列比对来估计一些边远的数据库中的匹配。

对于 PHI-BLAST，平常最常遇到的问题就是新用户对选择 PHI-BLAST 模式的规则没有感觉，而解决这个问题最好的方法就是用不同的信号序列多加练习。

5.9 网络资源

前文已经讲述了 3 种 BLAST 和相关的搜索工具：

① 进行 BLAST 搜索的专门的物种数据库（表 5-1）；

② 针对一些专门分子的 BLAST 站点（表 5-2）；

③ 包括 PSI-BLAST 在内的数据库搜索（表 5-5）的算法（表 5-3）。

表 5-5 提供 PSI-BLAST 的服务器

站点	网址
SBASE9.0(Trieste)	http://www3.icgeb.trieste.it/~sbasesrv/blast41n.html
NIH(日本)	http://spiral.genes.nig.ac.jp/homology/psi_blaste.shtml
巴斯德研究所	http://bioweb.pasteur.fr/seqanal/interfaces/psiblast.html
CMBI,荷兰	http://www.cmbi.kun.nl/bioinf/tools/psiblast.shtml
日本东京大学的人类基因组中心	http://blue1.ims.u-tokyo.ac.jp:8080/psiblast.html

5.10 问题讨论

1. BLAT 是一个相当快速和准确的程序，为什么它没有替代掉 BLAST 或者至少和 BLAST 一样得到普遍使用。

2. 在 PSI-BLAST 最初的应用过程中，算法进行多序列比对后将所有比对序列同源性≥98%的序列删除，只留下其中一条（Altschul 等，1997）。经过最近的改进后，程序判断的同源性为≥94%。如果将这个百分数调整为≥75%，你认为将会出现什么情况？你在实践中怎样证明你的想法？

5.11 问题

1. 创造一个人工蛋白质序列、人类 RBP4 蛋白质以及人类蛋白质激酶 C_α 的 C2 功能域。将这个联合序列输入进行 PSI-BLAST 搜索，结果如图 5-21 所示。一般来说，PSI-BLAST 程序能够搜索出多结构域蛋白质吗？自然界存在既有 lipocalin 又有 C2 域的蛋白质吗？

2. 疟疾寄生虫 *Plasmodium vivax* 有一个多基因家族 vir 为其所特有（Del Portillo 等，2001）。这些基因有 600～1000 个拷贝，他们可能会通过抗原变异导致慢性感染。选择 vir1 在非冗余数据库中进行 blastp 搜索，然后用同样的输入进行 PSI-BLAST 搜索。

(a) 在最初的搜索中，大约多少个蛋白质的 E 值小于 0.002，多少个 E 值大于 0.002？

(b) PSI-BLAST 搜索第一轮和第二轮循环之间增加的最好的新序列的分值是多少？

3. 我们以前用 HIV-1 pol（NP _ 057849）进行了一系列 BLAST 搜索。用这个蛋白质进行 blastp 搜索，观察分类报告中哪个病毒与之相匹配。然后，用 PSI-BLAST 重复这个搜索，将这次得到的分类报告与 blastp 搜索结果相比较，你观察到了什么？在 PSI-BLAST 搜索结果中有非病毒序列吗？你认为能够找得到吗？

4. 亨廷顿（Huntington）病是一种神经退行性失调。由于在人类亨廷顿基因中出现了 CAG 三核苷酸重复片段，从而导致了在人类亨廷顿基因中有大量的谷氨酸残基。人们对亨廷顿基因的功能并不是了解得很清楚，在一些真核生物中找到了它的同源物，但是在细菌中却没有发现过。如果能够在细菌中找到一个同源基因，将对研究亨廷顿基因的功能（同源蛋白的功能很可能相似）和蛋白质的进化有相当大的作用。

首先，用人类亨廷顿蛋白在非冗余数据库中进行一个标准的 blastp 搜索，得到的结果能够告诉你在这个数据库中是否存在一个明显的细菌同源蛋白质。然后进行 PSI-BLAST 和/或 PHI-BLAST 搜索来确定是否有细菌同源物。可以选择一个“信号序列”然后进行 PHI-BLAST 搜索，从而增加数据库中匹配的数目。注意这是一个在以往的文献中还没有给出答案的难题，还可能有其他很多种方法。

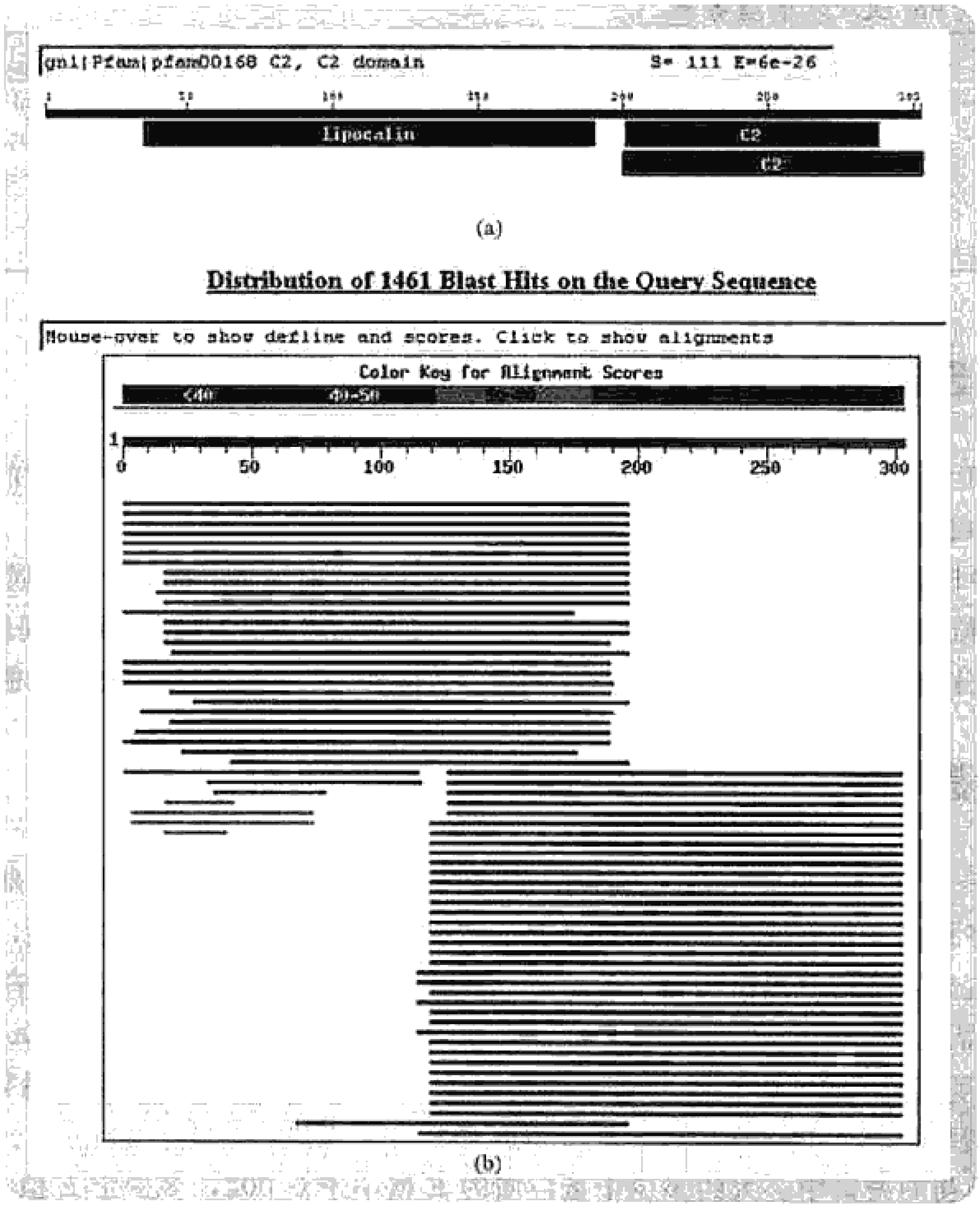


图 5-21 (a) 创建了一个由人类 RBP4 蛋白和蛋白激酶 C_a 组成的杂合蛋白质后，PSI-BLAST 格式页面显示两种功能域都被搜索出来了；(b) PSI-BLAST 的第三个步骤的结果表明这个算法有效地找到了和两个功能域都有关的数据库中的匹配。认为每个 lipocalin 蛋白和任何一个含有 C2 功能域的蛋白质是同源的这个观点是不对的。一般来说，所有的 PSI-BLAST 搜索结果列出了对应于一个 PSSM 的匹配

5.12 自测题

1. 在 EMBL 和 NCBI 数据库中未加工的 DNA 序列（与注释序列相比较而言）是：

(a) 完全重叠了；

(b) 很大程度上重叠了，不过序列不同；

(c) 相对只有一点点重叠

2. 当做下面哪个工作的时候，PSI-BLAST 搜索是最有效的：

(a) 在老鼠中找一个人类蛋白质的同源蛋白；

(b) 在数据库查询中找到更多的匹配蛋白质；

(c) 在数据库查询中找到更多的匹配DNA序列；

(d) 用模式序列或者信号序列加强数据库搜索

3. 下面哪一种 BLAST 程序是用氨基酸的信号序列在一个蛋白质家族中寻找匹配的？

(a) PSI-BLAST；

(b) PHI-BLAST；

(c) MS BLAST；

(d) WormBLAST

4. 下面哪一种 BLAST 程序用来分析免疫球蛋白最好？

(a) RPS-BLAST；

(b) PHI-BLAST；

(c) IgBLAST；

(d) ProDom

5. 在一个位点特异性打分矩阵中，列中可以有 20 个氨基酸，每一行则是查询的残基序列。在打分矩阵中，一个给定的氨基酸残基的分值是基于

(a) 一个 PAM 或者 BLOSUM 矩阵；

(b) 在多序列比对中出现的频率；

(c) 它出现的背景频率；

(d) 它周围的氨基酸的分值

6. 作为 PSI-BLAST 搜索过程的一部分，会给搜索序列和数据库中超过一定长度（例如 50 个氨基酸）的匹配进行打分。尽管这个比对的氨基酸残基序列没有改变，但是在接下来的 PSI-BLAST 循环中这个比对得到的分值可能会高一点或者低一点。这句话是否正确？

(a) 正确；

(b) 错误

7. 当一个位点特异性打分矩阵中混人了假阳性序列，就被称为“被破坏了”。下面哪一个选项是减少破坏的最好方法？

(a) 降低 E 值；

(b) 除去过滤；

(c) 用短一点的搜索序列；

(d) 减少循环次数

8. PSI-BLAST 和 PHI-BLAST 之间是什么关系？

(a) 它们都是在 NCBI 网页上发布的，但是两种没有联系的 BLAST 搜索；

(b) PHI-BLAST 搜索首先进行 PSI-BLAST 搜索，但是其匹配结果必须含有用户自己定义的短序列模式；

(c) 它们都采用位点特异性打分矩阵；

(d) 它们都进行连续的循环搜索

9. 如果想找一个和查询蛋白远源的蛋白质，下面哪种方法最可能成功？

(a) 采用 PHI-BLAST，因为你能自己选择一个和搜索蛋白质有关的信号序列；

(b) 采用 PSI-BLAST，因为这个算法使用位点特异性打分矩阵最为敏感；

(c) 采用 BLASTP，因为你能够调整你的打分矩阵从而使得搜索敏感度最大；

(d) 采用专门的物种数据库，因为它们中可能含有这种远源序列

10. 下面哪一个步骤能够使你相信找到的是一个新基因？

(a) 进行 PSI-BLAST 搜索；

(b) 搜索 EST 数据库看这个基因可能在哪里表达；

(c) 在 LocusLink 中看是否有这个基因家族中的其他成员已经被注释了；

(d) 在一个适当的数据库中用 BLAST 搜索你的新序列，以确认是否有其他人已经研究过你发现的新蛋白质

5.13 推荐读物

本章介绍了一些 BLASR 服务器。表 5-1～表 5-3 中所列大部分网站都有在线文档。PSI-BLAST 是由 Altschul 和他的同事们在一篇相当出色的文章中提出来的 (1997) (可

参见第 4 章的推荐读物）。PSI-BLAST 的进一步改进是由 Schaffer 等完成的（2001）。PHI-BLAST 算法则是由 Zhang 提出的（1998）。

参考文献

Altschul, S. F., et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

del Portillo, H. A., et al. A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. *Nature* **410**, 839-842 (2001).

Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8**, 967–974 (1998).

Friedberg, I., Kaplan, T., and Margalit, H. Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments. *Protein Sci.* **9**, 2278–2284 (2000).

Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).

Leonardo da Vinci. *Il Codice Atlantico*. Giunti, Florence, 2000.

Ning, Z., Cox, A. J., and Mullikin, J. C. SSAHA: A fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).

Park, J., et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201–1210 (1998).

Pearson, W. R., and Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448 (1988).

Schaffer, A. A., et al. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**, 2994–3005 (2001).

Wootton, J. C., and Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571 (1996).

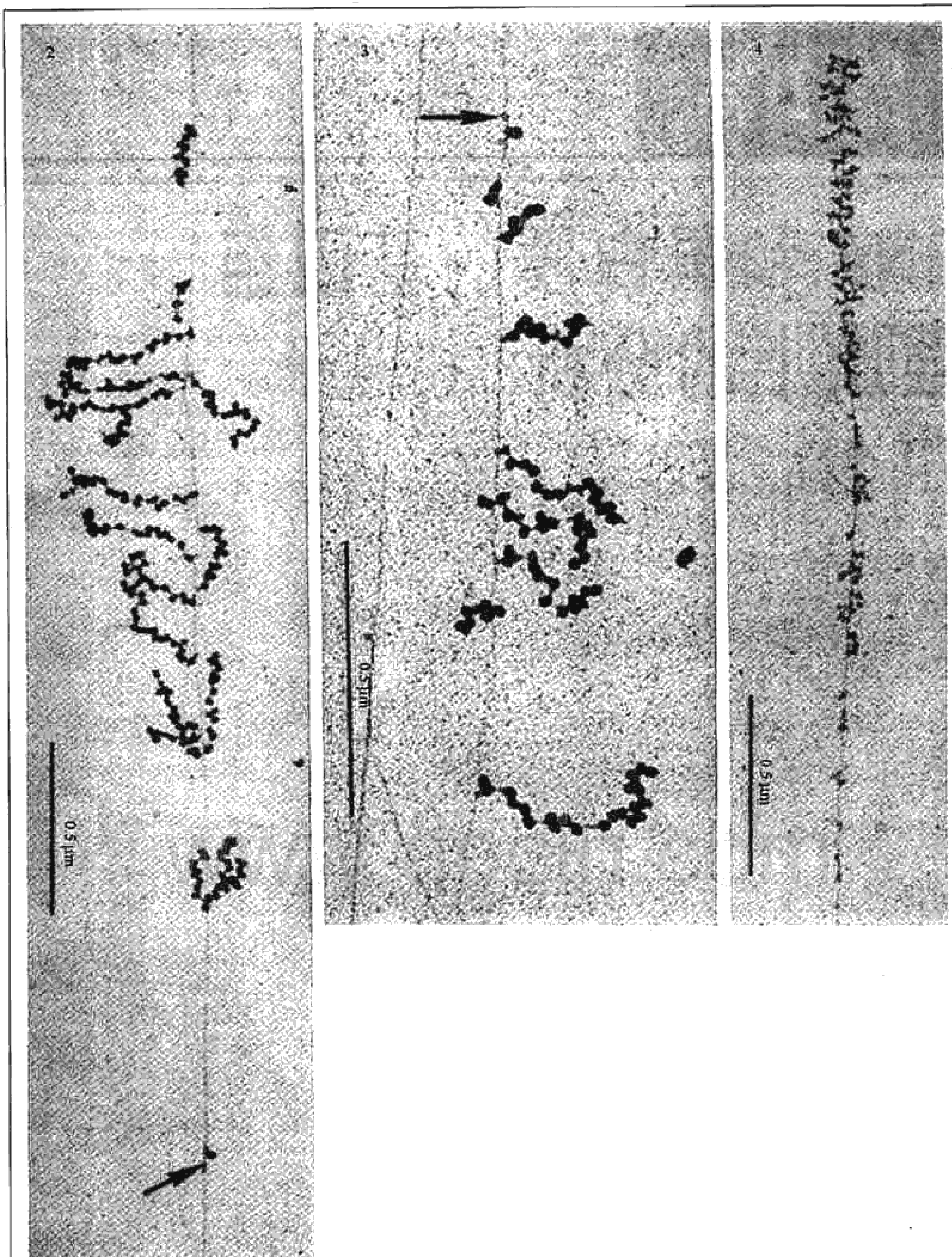
Zhang, Z., et al. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* **26**, 3986–3990 (1998).

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214 (2000).

第 2 篇

RNA 和蛋白质的基因组层次上的分析

第 6 章　基因表达的生物信息学方法
第 7 章　基因表达：芯片数据分析
第 8 章　蛋白质分析和蛋白质组学
第 9 章　蛋白质结构
第 10 章　多序列比对
第 11 章　分子水平的系统发生和进化



Miller 等（1970）将基因表达图像显现出来，他们显示出处于转录和翻译过程中的大肠杆菌染色体DNA（即每张图中垂直方向上的线状物）。当 mRNA 从基因组 DNA 转录出来并向染色体侧面延伸时，多核糖体（polyribosome）（黑色物体）像串起来的珠子一样出现在上面，正在将 mRNA 翻译成蛋白质。此图经许可采用

第6章

基因表达的生物信息学方法

6.1 引言

基因表达是在DNA转录为RNA的过程中发生的。每个真核细胞含有一个细胞核，其中带有5000～60000个编码蛋白质的基因。但是在任何给定情况下某细胞只将其基因中的一部分表达为mRNA转录物。基因组中表达的这一套基因一般被称为转录组。基因表达的调控有以下几种典型的基本方式：

① 组织部位的差别（如脑和肾）；

② 不同的发育阶段（如胚胎和成体组织）；

③ 对环境信号的动态响应［如被药物激活的即刻早期响应基因（immediate-early response gene）］；

④ 处于疾病状态；

⑤ 基因活性（如突变株和野生型细菌的比较）。

对各个物种中的基因表达谱的比较已经被用来研究大量的生物学问题（图6-1）。对于病毒和细菌，研究主要集中在病毒和细菌的基因表达以及宿主对病原入侵的响应上。对于真核生物，基因表达研究，尤其是微阵列（microarray）已经被用于研究一些基本问题，例如细胞周期或发育过程中基因激活情况的鉴别。人们已经研究了多细胞动物的细胞特异性基因表达，疾病在基因表达上的反映也已经在啮齿类和包括人类在内的灵长类动物中得到了研究。近年来，基因表达谱在基因组DNA序列的注释中起了很大的作用。对一个物种的基因组测序完成后，最重要的任务之一就是确定它编码了哪些基因（见第12章）。对表达基因的大规模测序，如对那些从cDNA文库（见本章后文）中分离出来的序列的测序，对于在基因组DNA中鉴别基因序列具有重要作用。

> 真核基因组基因含量的范围介绍见第16章。
>
> 除了将基因表达视为一个动态调控的过程，也可以认为在细胞中蛋白质和代谢物是被动态调控的，详见第8章。

近几十年来，已经运用很多技术对基因表达进行了研究，例如RNA印迹、反转录的聚合酶链式反应（RT-PCR）以及核糖核酸酶保护分析（RNase protection assay）。这些方法一次只能用来研究一个转录物。在RNA印迹法中，RNA经分离、凝胶电泳，并通过由单个基因得到的放射性cDNA探测。RT-PCR用特定的寡核苷酸对作为cDNA产物的特定转录

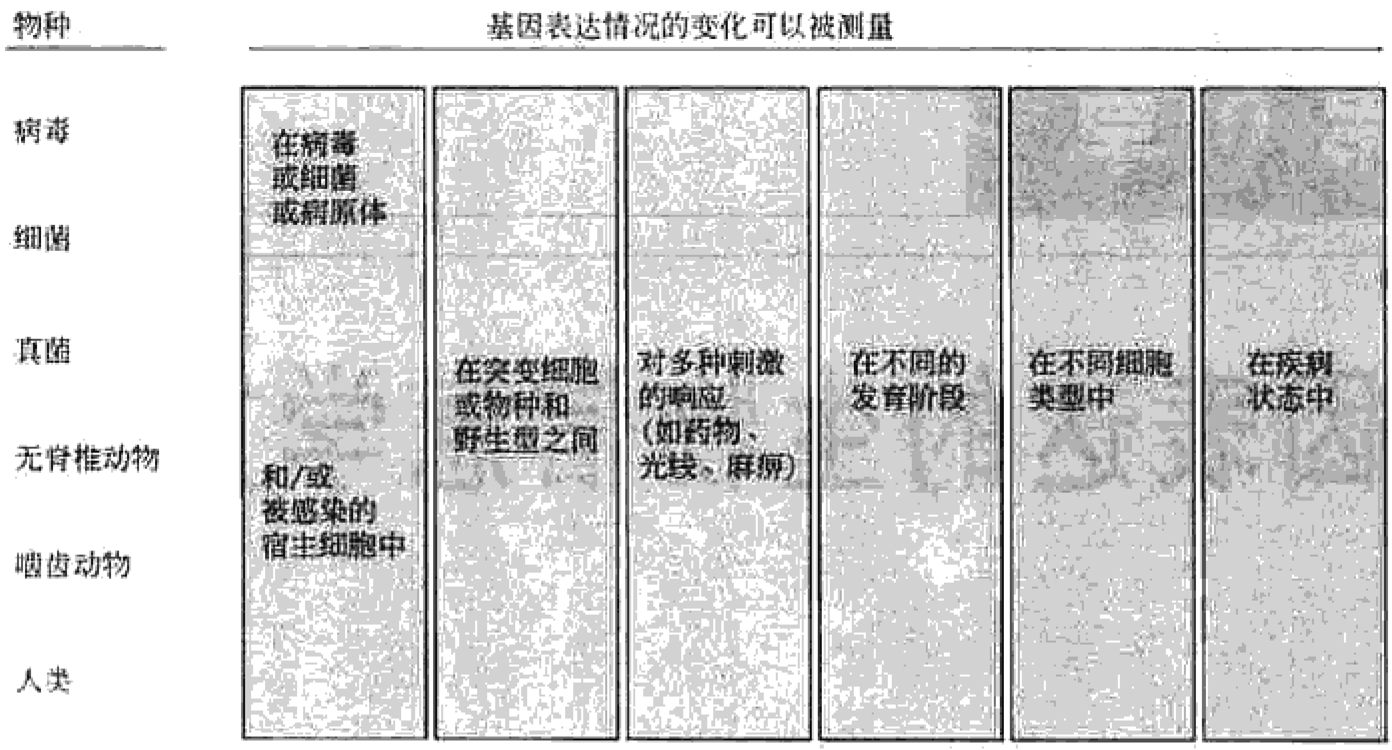


图 6-1 基因表达研究被用来处理各物种在不同条件下的生物问题。作为检测全基因组范围内基因表达变化的有力工具，高密度的 cDNA 或寡核苷酸微阵列已经出现，在本章和第 7 章中加以介绍。其他高通量技术被用来测量基因表达的变化，如表达序列标记（expressed sequence tags，EST）的分析和基因表达序列分析（serial analysis of gene expression，SAGE）

产物进行指数扩增。基于特定的体外转录 cDNA 能保护内源性转录产物使其不被核糖核酸酶降解的能力，可以用核糖核酸酶保护的方法对样本中的某个 RNA 转录产物进行定量测量。基因表达的比较能够在多种实验条件下进行（如正常组织和疾病组织基因表达的比较，经过药物处理和未经过药物处理的细胞系基因表达的比较）。典型的信号有荧光或放射性，故能被量化。信号归一化到大量管家基因或者预期表达水平保持不变的对照组。

出现了几种与这些方法不同的高通量技术，实现了对基因表达的大范围测量。基因表达的全局测量相对单基因表达的研究有以下两大优势。

① 大范围测量能够鉴别在一些生物状态下特异性表达的基因，如对各种人类组织基因表达的全局比较能够揭示哪些转录物是以区域性特异的方式表达的。虽然 RNA 印迹法一般不适用于在一项研究中测量大量基因的表达，但 Milner 和 Sutcliffe（1983）做了 191 个 RNA 印迹，他们发现其中 30%的基因在脑中有表达但在肝脏和肾脏中没有表达。这一做法现在能大尺度、更简单地通过高通量的技术实现。本章中将描述这些技术。

② 基因表达的高通量分析能够揭示生物样本中出现的基因表达模式和特征，例如蛋白产物功能相关的基因的协同表达。基因表达数据的分析工具（如聚类树）将在第 7 章介绍。

> 反转录酶常出现在反转录病毒中，是一种依赖于 RNA 的 DNA 聚合酶（即它从 RNA 转换得到 DNA）。

图 6-2 展示了几个基因表达的高通量研究方法。每种情况下两个（或多个）需要比较的生物样本中的总 RNA 或总 mRNA 被分离出来。RNA 一般通过反转录酶被转换成 cDNA。互补 DNA 和 RNA 相比，本质上更不容易被蛋白酶解或者化学降解，并且更容易被克隆、运输和测序。本章将探究以下 3 个基于计算机的基因表达分析方法：在 UniGene 中比较 cDNA 文库，在 NCBI 网站比较基因表达序列分析（SAGE）文库，以及 DNA 微阵列。

其他一些技术也能对基因表达进行测量，如差异显示技术（differential display）和消减杂交（subtractive hybridization）。虽然这些方法对于研究许多基因表达问题在技术上很成功，但

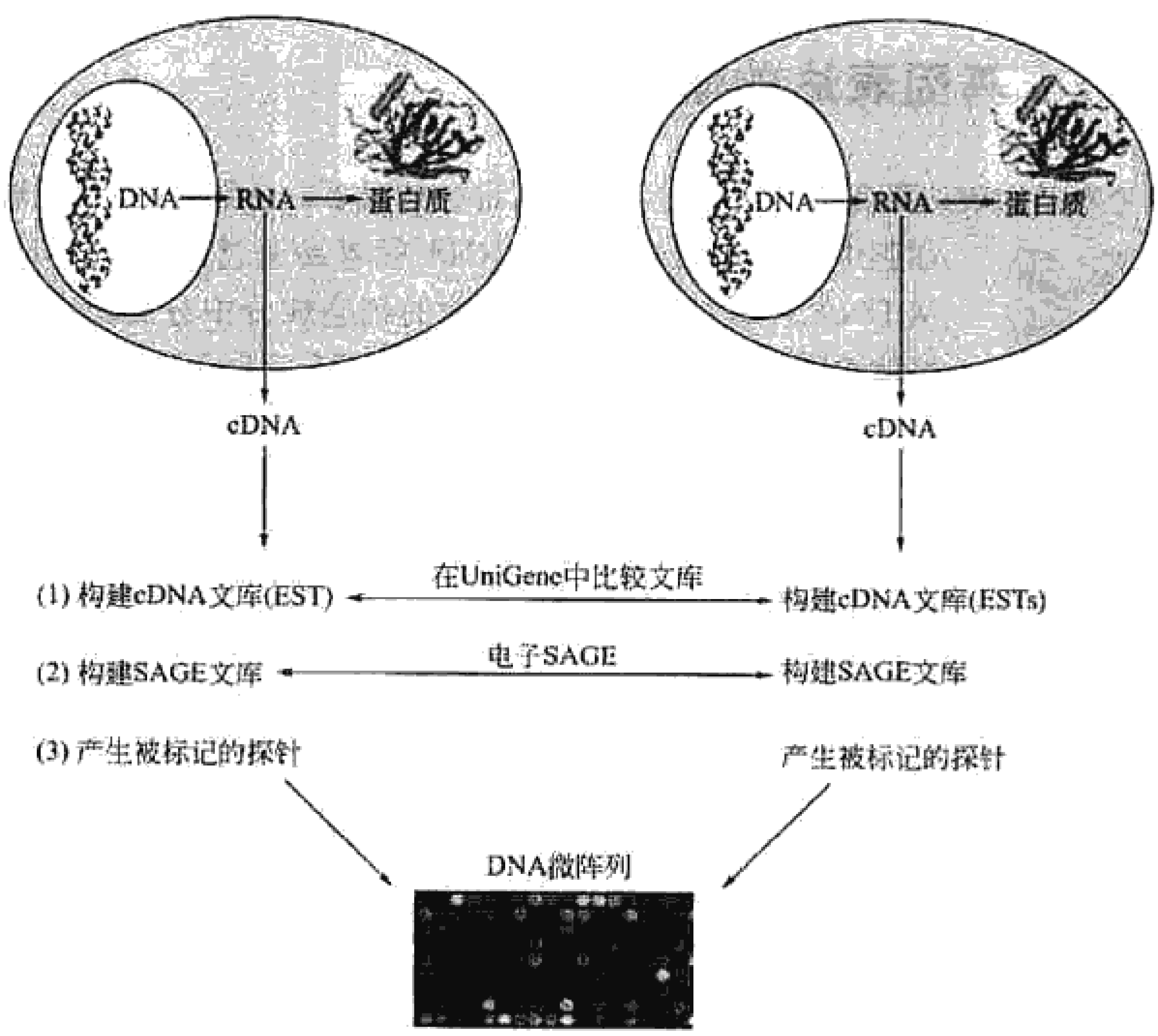


图 6-2　可以通过多种高通量技术测量基因表达。大部分情况下，对两个生物样本加以比较，如经过和未经药物处理的细胞系，受到和未受病毒感染的细胞，或老年和幼年大鼠的脑。RNA 可以被转换为 cDNA，从而可对细胞转录作更广泛的研究。本章和第 7 章我们将分析基因表达的三种研究方法：①通过产生表达序列标记，可以构建 cDNA 文库，并可在 UniGene 中进行电子比较；②基因表达序列分析是另一种可比较转录产物丰度的技术，也可以作电子方式的研究；③复杂的 cDNA 混合物可以被放射性或荧光标记，并在含有 cDNA 或对应于数千个基因的寡核苷酸的 DNA 微阵列上杂交

是它们和图 6-2 中显示的技术不同，并不涉及电子数据库的建立。基因表达测量技术的综述见 Watson 和 Margulies（1993）、Sagerstrom 等（1997）、Vietor 和 Huber（1997），以及 Carulli 等（1998）。

建立基因表达数据库后将其与 DNA 数据库比较是很重要的。DNA 数据库（如 GenBank）包含了 DNA 片段的序列信息，其大小范围为从小的克隆到整个染色体或者整个基因组。基因组 DNA 测序的出错率能够被测量（见第 12 章）。各实验室可以进一步验证 DNA 序列数据的质量。一般来说，某个物种的 DNA 序列是不随时间变化的，在不同的身体部位也一样。然而基因表达是依赖于环境的。一个基因表达数据库包含特定基因表达水平的定量测量。如果不同的实验室想刻画某细胞系的 RBP 表达水平，测量值会随很多因素变化，如细胞系的来源（如肝脏和肾脏）、细胞培养环境（如在亚融合和融合水平生长的细胞）、细胞环境（如生长介质的选择）、细胞的年龄、被研究的 RNA 的类型（分别含有不同数量杂质生物材料的总 RNA 或 mRNA）、测量技术，以及统计分析方法。因此，尽管建立类似 RefSeq 这样一个鉴别基因的高质量代表性 DNA 序列的计划是可能的，不过任何试图描述标准基因表达谱的尝试都必须考虑到转录发生时的大量因素。

6.2 mRNA：基因表达的研究对象

下面我们考察基因表达研究中的测量对象。绝大部分情况下，全部RNA都能够从感兴趣的细胞中被分离出来，这些RNA很容易通过离液剂（chaotropic agent）从DNA、蛋白质、脂质和其他的细胞成分中分离出来。总RNA由核糖体RNA（rRNA，占总RNA的70%以上）、转运RNA（tRNA，大于15%）和信使RNA（mRNA，约3%～4%）组成。蛋白质是由mRNA翻译而来的，其他种类的RNA将在第12章中介绍。

核糖体RNA分子构成了负责蛋白质合成的核糖体这一亚细胞单元的结构和功能成分。转运RNA分子携带一个特定的氨基酸分子，在蛋白质合成过程中与mRNA上对应的密码子匹配。转运RNA由70～90个核苷酸组成，形成典型的三叶草结构。下文会论及用于在真核生物DNA中寻找RNA基因的RNA序列数据库和软件（第16章）。

基因表达是通过一套复杂的步骤加以调控的，可以被分为以下4类：转录、RNA加工（RNA processing）、mRNA转运以及RNA监控（RNA surveillance）（Maniatis和Reed，2002）（图6-3）。

① 转录。基因组DNA通过一些高度调控的步骤被转录为RNA。20世纪70年代，基因组DNA的序列分析显示DNA的某些部分（被称为外显子）和相应的mRNA邻近的开放阅读框相匹配，而基因组DNA的其他区域（内含子）代表成熟mRNA中并不出现的插入序列。

② RNA加工。剪接体将内含子从前体mRNA中切除。剪接体由5个小核RNA（snRNA）的复合体和超过70个蛋白质分子组成。当剪接体选择性地纳入或排除特定的外显子时，发生可变剪切（Modrek和Lee，2002），并且前体mRNA被在5′端加帽（真核细胞mRNA含有称为帽子的反向鸟嘌呤结构）。成熟mRNA的3′端

Richard J. Roberts和Phillip A. Sharp因发现“断裂基因”获得1993年的诺贝尔生理学和医学奖。详见http://www.nobel.se/medicine/laureates/1993。

果蝇的一种分子提供了可变剪切的一个很好的例子。唐氏综合征细胞粘连分子（Down syndrome cell adhesion molecule，DSCAM）的基因产物含有38000种以上可能的异型（isoform）（Schmucker等，2000；Celotto和Graveley，2001）。该基因包含95个组织成簇状的可选择外显子。在功能上，多种DSCAM蛋白可赋予果蝇的神经元连接以特异性。

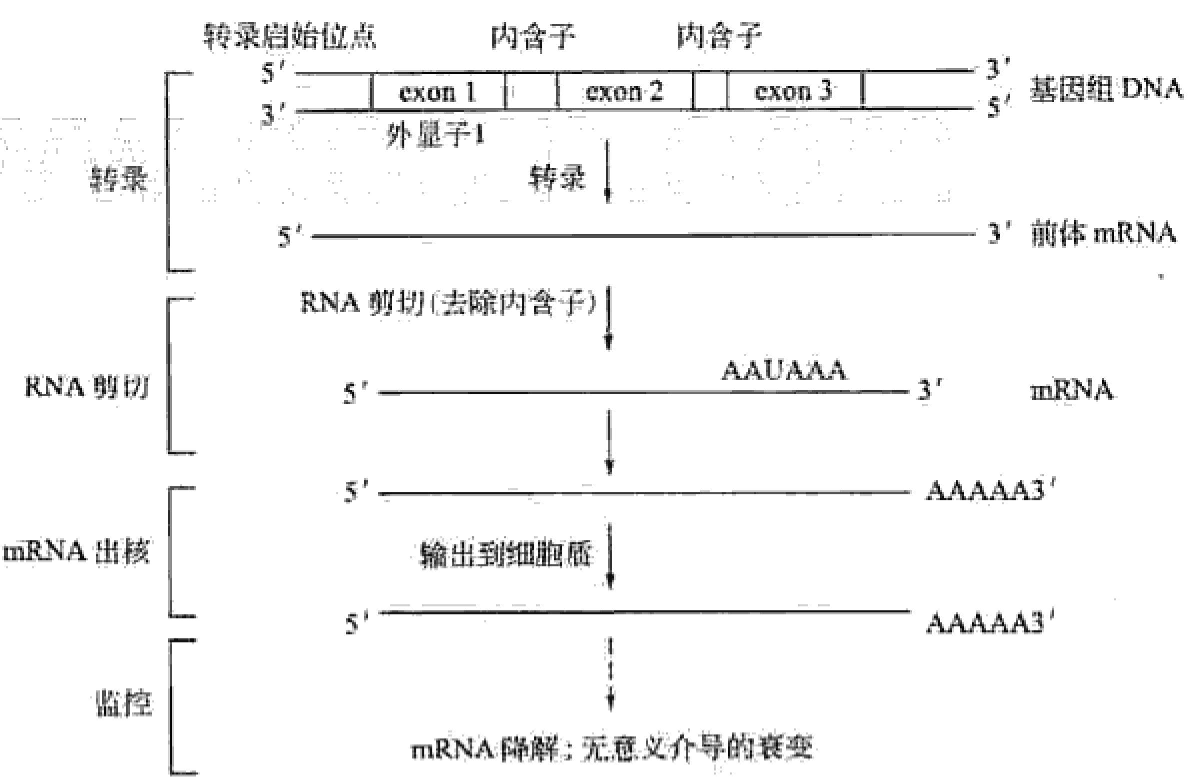


图6-3 真核基因的RNA处理。基因组DNA包含外显子（对应于成熟mRNA）和内含子（插入序列）。DNA被转录后，前体mRNA在5′端加帽，剪切去除内含子。多聚腺苷酸化信号（一般为AAUAAA）被识别后，RNA被核酸内切酶在下游10～35个核苷酸处截断，并加上polyA尾巴（一般长100～300个残基）。多聚腺苷酸化的mRNA被输出到细胞质并在核糖体上翻译成蛋白质。RNA监控系统涉及通过无意义介导的衰变来降解异常mRNA。图中的虚线指RNA监控机制也可以降解前体mRNA

有一长串腺嘌呤残基，这一特点在所有核酸中很独特，在其上游 10～35 个核苷酸处有多聚腺苷酸化信号 AAUAAA 或 AUUAAA。mRNA 的多聚腺苷酸化对于实验十分方便，因为寡核苷酸（由一串附着于固体支持物［oligo(dT)resin］上的胸腺嘧啶组成）能被用于快速分离高纯度的 mRNA。有时基因表达研究使用总 RNA，很多其他情况下则用 mRNA。

③ RNA 出核。剪接完成后，RNA 从细胞核输出到细胞质。翻译将在细胞质中进行。

④ RNA 监控。真核生物通过许多 RNA 监控方法监视前体 mRNA 和 mRNA 分子可能发生的无义突变（产生不合适的终止密码子）以及移码突变（Maquat，2002）。这个因为产生无意义密码子而产生的衰变机制（nonsense-mediated decay mechanism）对于维护功能性 mRNA 分子很重要。别的机制控制 mRNA 的半衰期，将其作为降解的目标从而调控其有效性。

图 6-4 显示了一个 mRNA 序列如何与基因组 DNA 比对的实际例子。访问 LocusLink

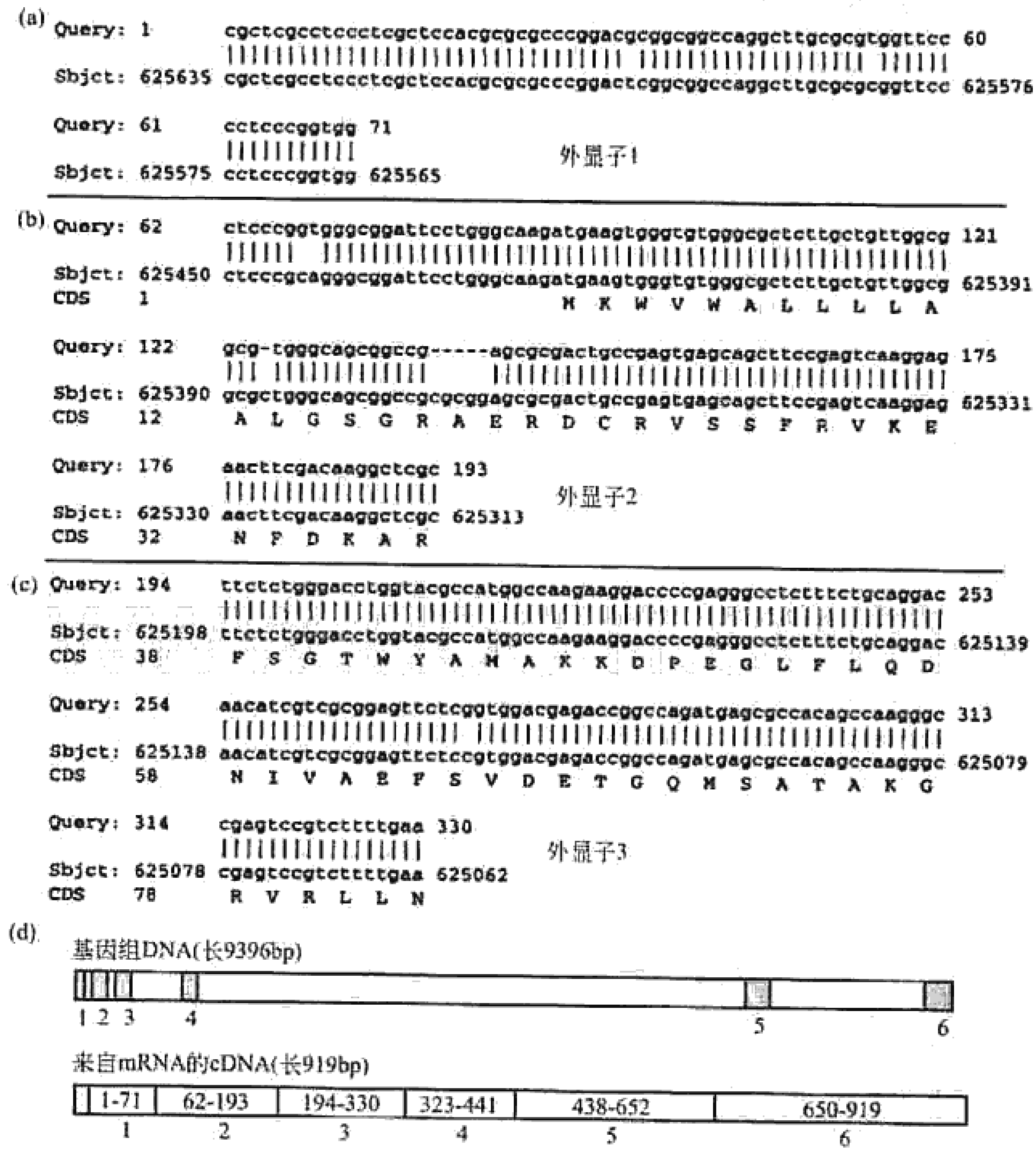


图 6-4 通过由 mRNA 得到的人类 *RBP4* 基因序列（来自 LocusLink；NM_006744）和包含人类 *RBP4* 基因的基因组序列（也来自 LocusLink：NT_030059）之间的序列比对，可以证实 mRNA 和基因组 DNA 之间的关系。(a)～(c) 显示了 BLAST2 比对的一部分。被称为重叠群（contig）的基因组 DNA 片段由 10 号染色体上的 20700699bp 组成，沿着下面那条单链的方向和 *RBP4* 比对（所以“subject”行的数字递减）。如果查询序列和被内含子隔断的基因组 DNA 序列匹配，就能明白其含义。(d) 显示了基于该比对的基因组 DNA（上面）和连续 mRNA（下面），包括两两比对的位置，外显子用 1～6 标记。注意其中的标度不同

网站并选择*RBP4* mRNA的序列号（即对应于该mRNA的DNA序列，NM _ 006744）和人类10号染色体上包含*RBP4*基因的一段长度超过2×10^7bp（残基）的基因组DNA片段（NT _ 030059）。用BLAST2进行的序列比对显示两者匹配得很好，其中大段的间隙对应于基因组DNA中的内含子。

可以注意到图6-4中的比对有空隙、错配，或两者都有。这些不一致反映了多态性或者基因组DNA测序或cDNA测序的差错。判断错误出在哪个序列的方法之一是寻找一致性：如果多个单独得到的基因组DNA克隆或EST在我们感兴趣的区域内有一样的核苷酸序列，就更能确信序列的正确性。对此第12章有进一步的讨论。

6.3 cDNA文库的基因表达分析

对于感兴趣的组织样本或生物系统中表达的多数mRNA分子如何进行研究呢？技术上最直接的方式是从少量组织样本中适度分离出大量的RNA或mRNA（比如数百微克的总RNA）。然而RNA并不稳定且高度复杂，往往包含数千种不同的转录产物。解决这个问题的方式之一是生成一个cDNA文库（图6-5）。简单地说，就是将RNA转变成双链cDNA，克隆到一个载体中，然后在一个细菌细胞系中增殖。载体（如质粒）具有尺寸小、生长快的特点，并且有能力容纳来自初始组织样本或其他生物来源的单个cDNA插入片段。数以千计的cDNA文库已经被商品化，每个文库都是从特定的物种、细胞类型、发育阶段和生理条件下得到的。cDNA文库的克隆可被移植到皮氏培养皿（Petri dish）

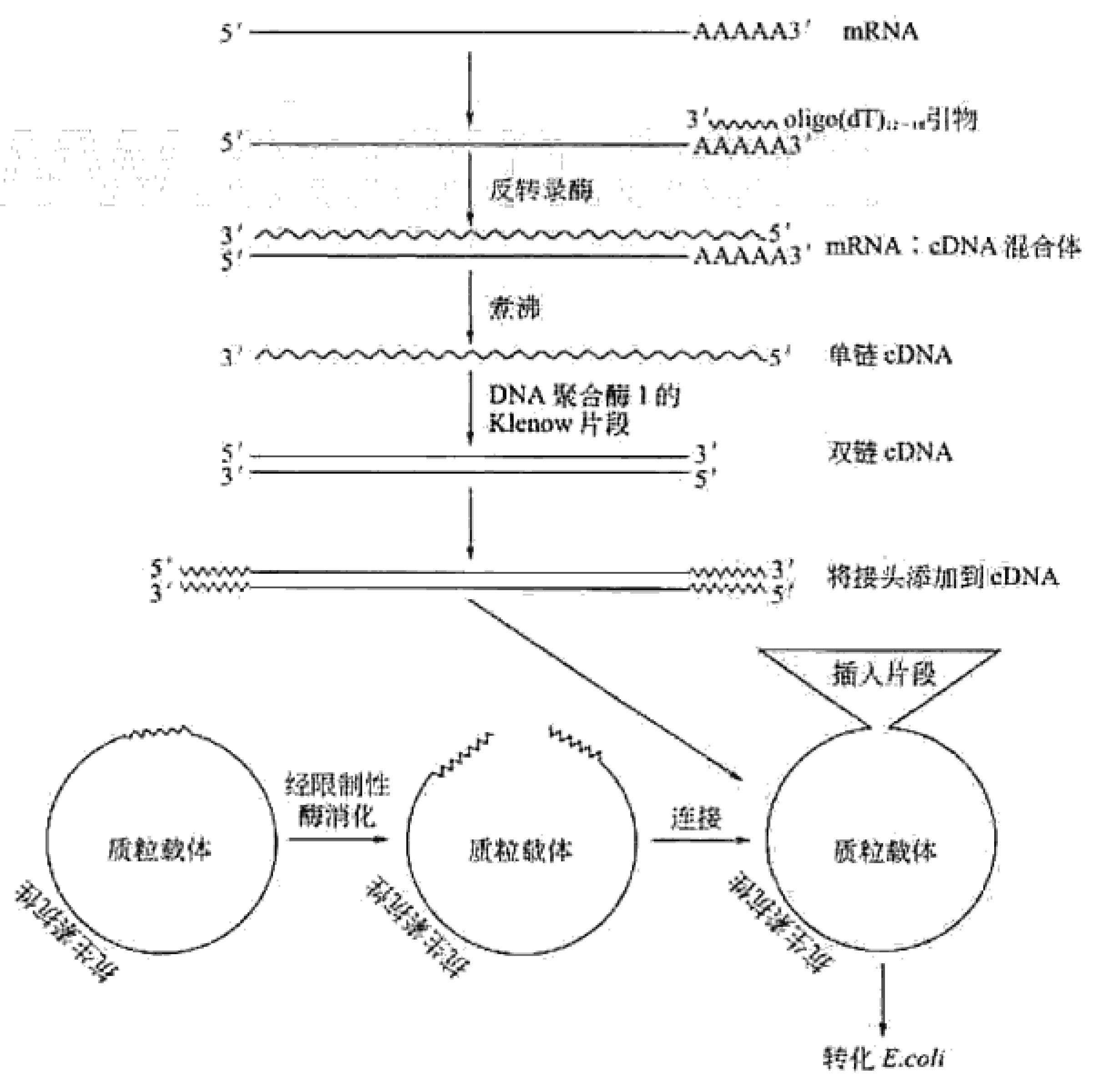


图6-5 cDNA文库的构建。信使RNA在3′端和oligo(dT)引物杂交，经过反转录得到一个mRNA：cDNA混合体。煮沸使RNA变性后，用DNA聚合酶合成cDNA双链。将接头（linker）（即可被限制性内切酶识别的核苷酸序列）加到cDNA上，于是对该cDNA（又称为插入片段）和质粒或者噬菌体（又称为载体）一起进行合适的消化后，两者可以被连接起来。然后将大肠杆菌转化并经抗生素抗性筛选，即可形成cDNA文库，其中含有来自初始mRNA样本的数千个独立的cDNA插入片段

上。然后可以对这些 cDNA 插入片段，即对表达序列标记进行测序。

> 对 GenBank 中 EST 的数量的归纳见 http://www.ncbi.nlm.nih.gov/dbEST/dbEST-summary.html。

被测序的 EST 总数已达数百万，通常是 cDNA 克隆 3′端或 5′端长约 500bp 的单股读取（single-pass read）片段。Adams 等（1991～1993）率先对数千个 EST 进行测序，以鉴别特定组织中被表达的基因（这就是所谓的鸟枪法单股读取）。这些研究揭示了在最高相对水平上哪些基因被表达（如人脑中的β-肌动蛋白和髓磷脂基质蛋白），以及脑部不同区域的基因表达变化。

> UniGene 可通过以下网址访问：http://www.ncbi.nlm.nih.gov/UniGene/。

前面介绍的 UniGene 是一个将 EST 分成无冗余的簇的集合的系统（见第 2 章）。原则上，每个单独的基因（"UniGene"）都被赋予一个单独的 UniGene 条目。UniGene 同时包括特性已经比较清楚的基因和仅由 EST 推测其存在的基因。所有对应于基因的 EST 都有一个特定的 UniGene 序列号。含有和某个已知基因相似的 EST 的 UniGene 簇被归为和该基因"高度相似"（比对区域的一致性超过 90%），"一般相似"（一致性介于 70%～90%）或"弱相似"（一致性小于 70%）。

表 6-1 UniGene 的"Human"条目下簇大小的直方图（Build 153，*Homo sapiens*）

簇大小	簇数量
1	36206
2	14384
3～4	15804
5～8	10612
9～16	5852
17～32	3986
33～64	3516
65～128	4095
129～256	3953
257～512	2170
513～1024	709
1025～2048	213
2049～4096	70
4097～8192	26
8193～16384	5
16385～32768	1

注：来源于 http://www.ncbi.nlm.nih.gov/UniGene/Hs.stats.shtml。

虽然两个 UniGene 簇可能合并成一个，人类 UniGene 簇的数目可以作为对人类基因数目的一个估计。每个簇有若干序列，其数量从 1（singleton）到 10000 个以上（表 6-1）。表 6-1 的 104170 个簇中，1/3 以上是 singleton，这表明这些基因很少被表达，以致只被观测到一次。另一种可能是，它们并不是可信的表达基因，而是来自文库的污染。UniGene 簇的数目超过 10 万个这一情况的确和对人类基因组中编码蛋白质的基因数量在 35000 个左右的估计（见第 17 章）不一致。Singleton 也可能代表了尚未被测序的那部分基因（见第 2 章）。

数以千计的 UniGene 簇很小，这表明有些基因很少表达。其他基因（如肌动蛋白和微管蛋白）则被频繁地表达。甚至有些并不对应任何已知的得到注释的基因的 EST 簇却频繁表达。表 6-2 中显示了 UniGene 所有物种中最大的那些簇的情况。

20 世纪 70 年代进行的 DNA 杂交实验显示，一个典型的哺乳动物细胞表达来自 10000～30000 个不同基因的约 300000 个 mRNA 转录产物分子（Hastie 和 Bishop，1976）。Bishop 与其合作者（1974）根据相对丰度将 mRNA 分成以下 3 类：①以高丰度表达的基因（占总转录产物约 10%）；②中等丰度表达的基因（占 45%）；③低丰度基因（占 45%）。这 3 类基因可对应于 UniGene 中不同大小的簇（表 6-1），不过表中簇大小并没有准确标出。

> 与 UniGene 中肌动蛋白这一项（Hs.1288）对应的 EST 有 813 个，与α-微管蛋白（Hs.334842）相关的序列则有 7237 个（2002 年 10 月的数据）。

由于所有 EST 都是在体内某区域的特定发育阶段和生理状态下得到的，对 EST 的分析是和大量内在信息相关联的（Schuler，1997）。从 UniGene 中获取信息的方法主要有两种。第一种方法是，如果想知道某个基因（如 RBP）在体内什么地方表达，可以

表 6-2 UniGene 所有物种中最大的簇（2002 年 10 月）

UniGene 编号	物种	簇大小	基因名
Hs. 181165	人(*H. sapiens*)	25232	真核翻译延伸因子 1α 1
Hs. 169476	人(*H. sapiens*)	14277	甘油醛-3-磷酸脱氢酶
Ta. 9227	小麦(*Triticum aestivum*)	14231	泛素
Hs. 14376	人(*H. sapiens*)	12749	肌动蛋白 γ 1
Mm. 196614	小鼠(*M. musculus*)	10649	真核翻译延伸因子 1α 1
Hs. 356360	人(*H. sapiens*)	10596	核糖体蛋白 S2
Mm. 30266	小鼠(*M. musculus*)	10290	血红蛋白，成体 β 主链
Hs. 356428	人(*H. sapiens*)	9987	仅胎盘绒毛中表达的 mRNA
Hs. 288061	人(*H. sapiens*)	9667	肌动蛋白 β
Dr. 2984	斑马鱼(*Danio rerio*)	9058	40S 核糖体蛋白 S18

注：获取以下数据的方式是，先通过文件传输协议（FTP）下载所有 UniGene 文件，并在关系数据库（DRAGON）中按大小分类（Bouton 和 Pevsner，2000）。

搜索整个 UniGene 库，与该基因相关的 EST 数量反映了该转录产物的丰富程度（并见下文的讨论），而 EST 的来源文库则反映该 EST 的区域分布情况。这个方法有时被称为电子 RNA 印迹法（electronic Northern blot）。从 UniGene 提取信息的另一种方法是 cDNA 文库的电子扣除。UniGene 中的 cDNA 文库电子扣除（electronic cDNA library subtraction）允许用数字差别显示（digital differential display）工具对两个样本中的 cDNA 序列进行比较［图 6-6(a)］。

实验上，差别显示技术对两个 RNA（或对应的 cDNA）样本相互扣除，一个样本被选择性标记（如将一个寡核苷酸序列连接到一个 cDNA 群体上），然后两个样本杂交，在共有的克隆之间形成二联体，接着单个样本中大量多出的 cDNA 克隆被选择性扩增并测序。

从 UniGene 主页可以使用这一工具。首先选择一个物种（如人类），然后点击“Digital Differential Display”［图 6-6(b)］。对肌肉文库和脑文库的比较显示出大量转录产物选择性地和肌肉关联（如肌红蛋白和肌钙蛋白）（图 6-7），其他转录产物则选择性地和脑关联。输出结果中每个基因有一个点，点的密度对应其表达水平。用“Fisher's exact test”对每个转录产物可计算得到一个相应的概率，用来检验以下零假设：任何给定基因的序列数量在两个库（如肌球蛋白）中是相等的（表 6-3）。

表 6-3 Fisher's 2×2 Exact Test 用来检验零假设［即某个给定的基因（基因 1）在两个群体中没有差异调控］

项目	基因 1	所有其他基因	总和
群体 A(如脑)	归属基因 1 的序列数($g1_A$)	该群体中非基因 1 的序列数(N_A-g1_A)	N_A
群体 B(如肌肉)	归属基因 2 的序列数($g1_B$)	该群体中非基因 1 的序列数(N_B-g1_B)	N_B
总和	$c=g1_A+g1_B$	$C=(N_A-g1_A)+(N_B-g1_B)$	

注：来源于 Claverie（1999）和 http://www.ncbi.nlm.nih.gov/UniGene/fisher.shtml。

Fisher's exact test 的 p 值由下式给出：

$$p=\frac{N_A!N_B!c!C!}{(N_A+N_B)!g1_A!g1_B!(N_A-g1_A!)(N_B-g1_B!)} \tag{6-1}$$

如果概率值 p 小于 $0.05/G$，零假设（基因 1 在大脑和肌肉中的调控没有差别）不成立。

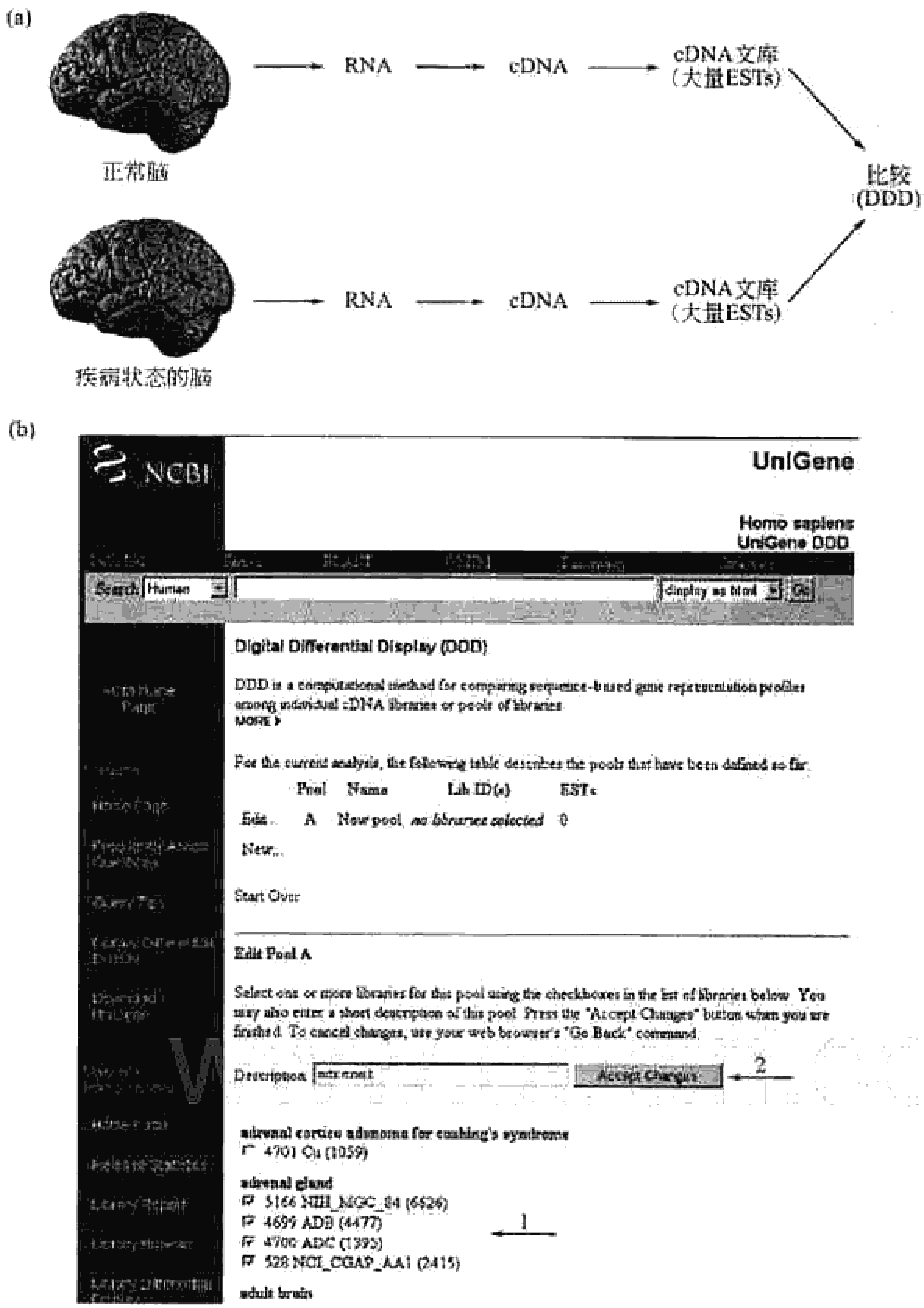


图 6-6　数字差别显示（DDD）被用来比较 UniGene cDNA 文库中表达序列标记的内容。已经建立了 1000 多个文库，其步骤是先从不同组织来源［如正常状态和疾病状态的人脑（a）］中分离出 RNA 并合成 cDNA，然后包装 cDNA 文库中的这些 cDNA，并对各文库中数千个 cDNA 克隆（EST）测序。（b）各文库（或多个文库组成的集合）中的克隆可用 DDD 进行比较。这里显示的网站可以从 NCBI 的 UniGene 站点进入，步骤是在左侧工具条点击“Homo Sapiens”，然后选择“Library digital differential display”。在本网页点击对应于任何文库（或一组文库；箭头 1）的方框，然后点击“Accept changes”（箭头 2）。接下来就可以选择第二个（或第二组）文库进行比较

这里 0.05 是显著性的阈值，G 是被分析的 UniGene 簇的个数（所以 G 是不变的 Bonferroni 修正）。

用各种文库构成的数据库进行基因表达谱比较可以作为一个工具，以快速提供候选基因作进一步分析。在进行的肌肉和大脑的文库比较中，通过数字差别显示得到的被调控的基因中有一些是功能未知的“假定蛋白”，有待进一步研究。Schmitt 与其同事（1999）用类似的方法在乳腺癌组织中找到 139 个被选择性向上调节的转录产物。这些转录产物可作为乳腺癌早期探测的标记，也可能反映肿瘤组织的变化，并对治疗介入提供靶子。大量其他研究中则运用了 EST 测序以及对已测序的 cDNA 文库的电子分析等手段（如 Carulli 等，1998）。

除了 NCBI 应用的 Fisher's exact test 以外，还有其他的统计方法用于 cDNA 文库的比较，特别是 Stekel 等人（2000）发展了一个对数似然方法，用来估计在两个或多个 cDNA 文库的比较中，观测到的基因表达差别由真正的转录差别而非取样误差造成的概率。

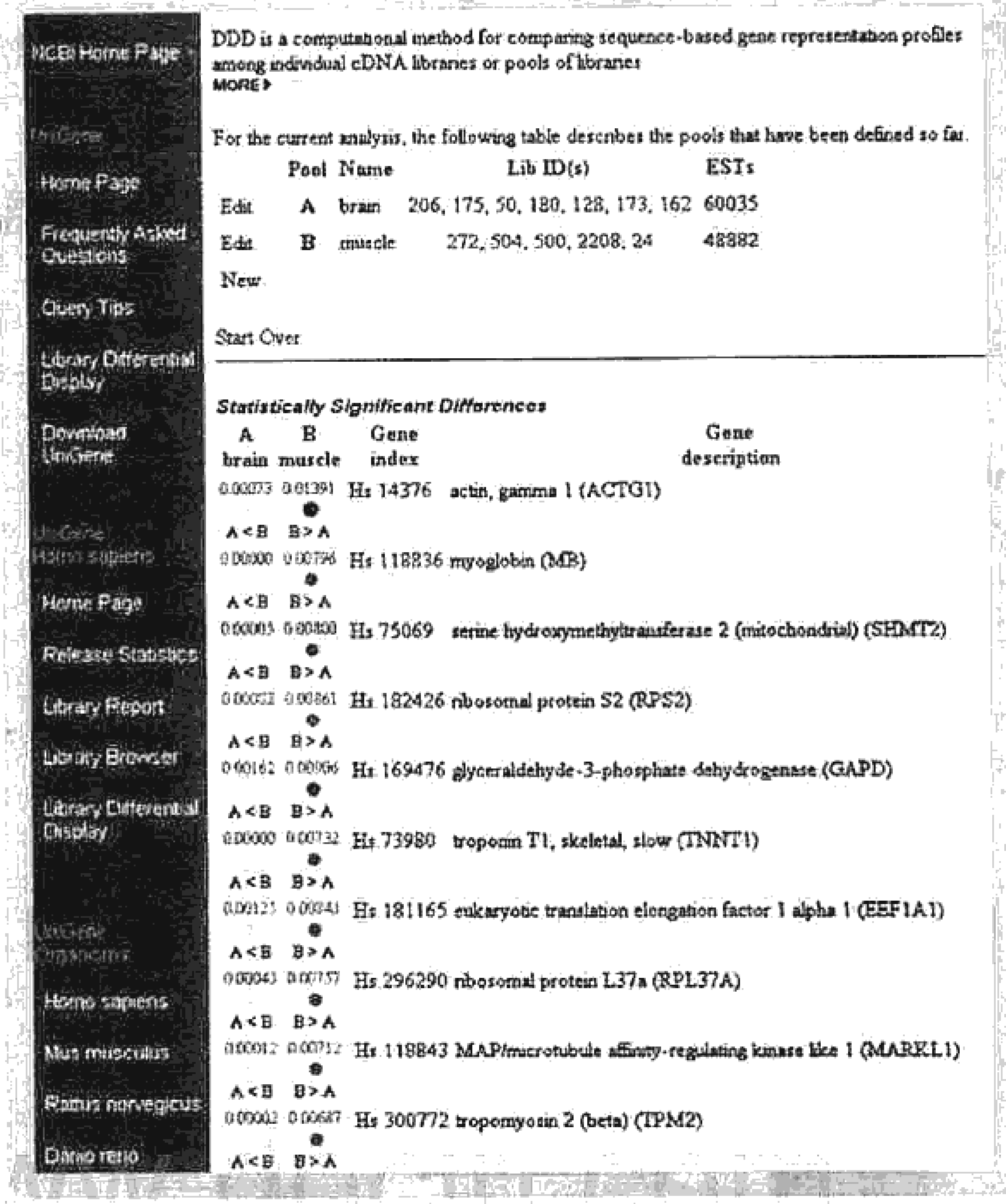


图 6-7 使用 UniGene 的 DDD 工具得到的 cDNA 文库电子比较结果。结果显示为一列基因（包括其 UniGene 序列号），这些基因在其中一个或一组文库中表达更多。这里显示的是多种肌肉特异性的转录产物（如肌钙蛋白、原肌球蛋白），其他转录产物（如成神经细胞瘤基因，未显示）则在脑文库中表达量更高

6.3.1 用 cDNA 文库解释表达数据的注意事项

在分析 UniGene 和其他来源的 cDNA 文库内容时必须很谨慎，原因如下。

① 选择建立什么文库由研究者决定，所以可能会较多偏向较熟悉的组织（如脑和肝脏），而较少偏向不寻常的组织。大鼠的鼻子含有至少 10 多个分泌腺，几乎所有这些分泌腺的功能都是未知的，而大部分这些腺体的 cDNA 文库尚未建立。

在 UniGene 网站，点击"Homo sapiens"，然后点击"library browser"，就可以看到在典型的文库中被测序的克隆涵盖的范围。

② 一个文库被测序的程度影响其对原细胞或组织内容的代表能力。cDNA 文库中克隆的出现次数被认为应该忠实地反映原样本中转录物的丰度。如果仅有 500 个克隆被测序，分析整个文库的内容时大量低丰度的转录物将不大可能得到反映。实际上各 cDNA 文库的测序程度是不同的。

③ 有一种偏离来源文库的构建。很多文库有一个归一化的过程，这个过程中频繁表达地转录产物的出现次数减少，而很少表达的转录产物被表现得更频繁。文库归一化的目的是尽量减少对高度表达的基因不必要的重复测序，从而发现罕见的转录产物（Bonaldo 等，1996）。直接用 UniGene 的差别显示工具在已归一化和未归一化的文库之间进行比较是不合适的。

④ EST 一般仅在双链中的一条链上被测序，而不对两条单链都测序。所以相对于被完

全测序的序列，出错率更高（测序的出错率将在第 12 章中讨论）。

⑤ 嵌合序列可能污染 cDNA 文库。例如，构建文库时两条不相关的插入序列偶尔会被克隆到同一个载体中。

6.3.2 TIGR 基因索引

针对 EST 和 cDNA 文库除了前面讨论的 UniGene 外，还有其他重要资源。基因组研究所（the Institute for Genomic Research Institute）提供的 TIGR 基因索引（gene indice），是按数十个物种特异性数据库组织起来的一套 EST 数据（Quackenbush 等，2001；Lee 等，2002）。TIGR 对 EST 序列的分析方法注重于将其聚集成所谓假设一致性（tentative consensus，TC）序列的具有惟一性的基因。通过强调聚类和集合，生成和基因对应的一套一致性序列。对于用 UniGene 数据不容易处理的问题，TIGR 基因索引可能更合适（UniGene 并没有将 EST 聚集成单一的簇，而 NCBI 对每个簇提供最长的 EST 的序列）。TIGR 真核基因直系同源（eukaryotic gene orthologs，EGO）数据库由通过 TC 序列的两两比对得到的直系同源基因组成，可以对数十种生物中的同源基因进行比较。TIGR 基因索引也可以通过 blastn 和 tblastn（表 5-1 和图 5-3～图 5-5）访问。

从 IMAGE 协会网站（http://image.llnl.gov/）可以查询许多物种的克隆信息。

EST 可以通过几种方式和全长克隆联系起来。基因组及其表达的集成分子分析（integrated molecular analysis of genomes and their expression，IMAGE）协会搜集了一些文库并使其可以分发使用。哺乳动物基因集成（MGC）是一个旨在提供人类和小鼠全长克隆的 NIH 计划，其网站（图 6-8）可通过 BLAST 搜索。

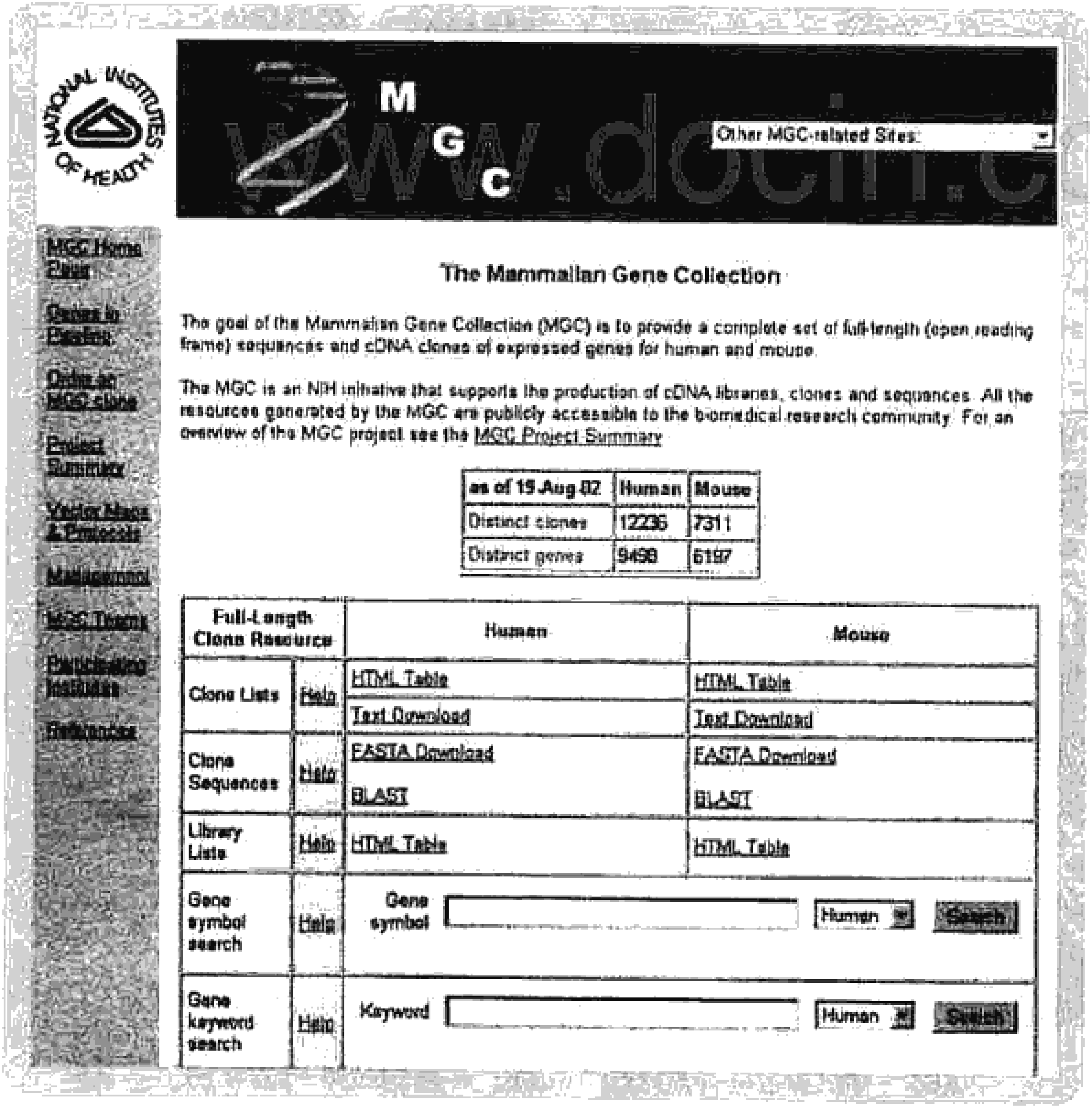


图 6-8 哺乳动物基因集成（mammalian gene collection，MGC）是一个对人类和小鼠所有基因被表达的开放阅读框克隆进行测序的计划。目前已经收集了 20000 多个人类和小鼠的无冗余克隆，其中 3000 多个被完全测序，这些被选择的克隆中 75% 为全长开放阅读框。MGC 是由 NIH 发起的，它将为这一领域提供可靠的高质量克隆资源，其网站（http://mgc.nci.nih.gov/）允许 BLAST 搜索

6.4 基因表达系列分析

基因表达系列分析（SAGE）通过测量感兴趣的组织中大量转录产物，实现对基因表达的定量测量。从转录产物3′端分离出长9～11bp的DNA序列作为标签（tag），测序后赋予该基因。

NCBI 的 SAGE 网站为 http://www.ncbi.nlm.nih.gov/SAGE。

图6-9说明了产生SAGE标签的过程（Velculescu等，1995）。先从感兴趣的样本中分离出RNA，并用生物素化的寡聚脱氧胸苷引物转化成cDNA。然后用限制性酶（“锚定酶”，anchoring enzyme）消化总体转录产物，消化后分离得到一些短片段，其中由于生物素和抗生物素蛋白之间的紧密作用，使得转录产物的3′端被束缚在链霉素抗生物素蛋白的微珠上。接着将两种接头各自添加到cDNA上，经过特殊的限制性酶切释放出带有接头的cDNA短片段（即“标签”）。最后将这些标签连接起来，克隆后测序。经过整个过程从生物来源中产生数千个（甚至数百万个）标签的记录。

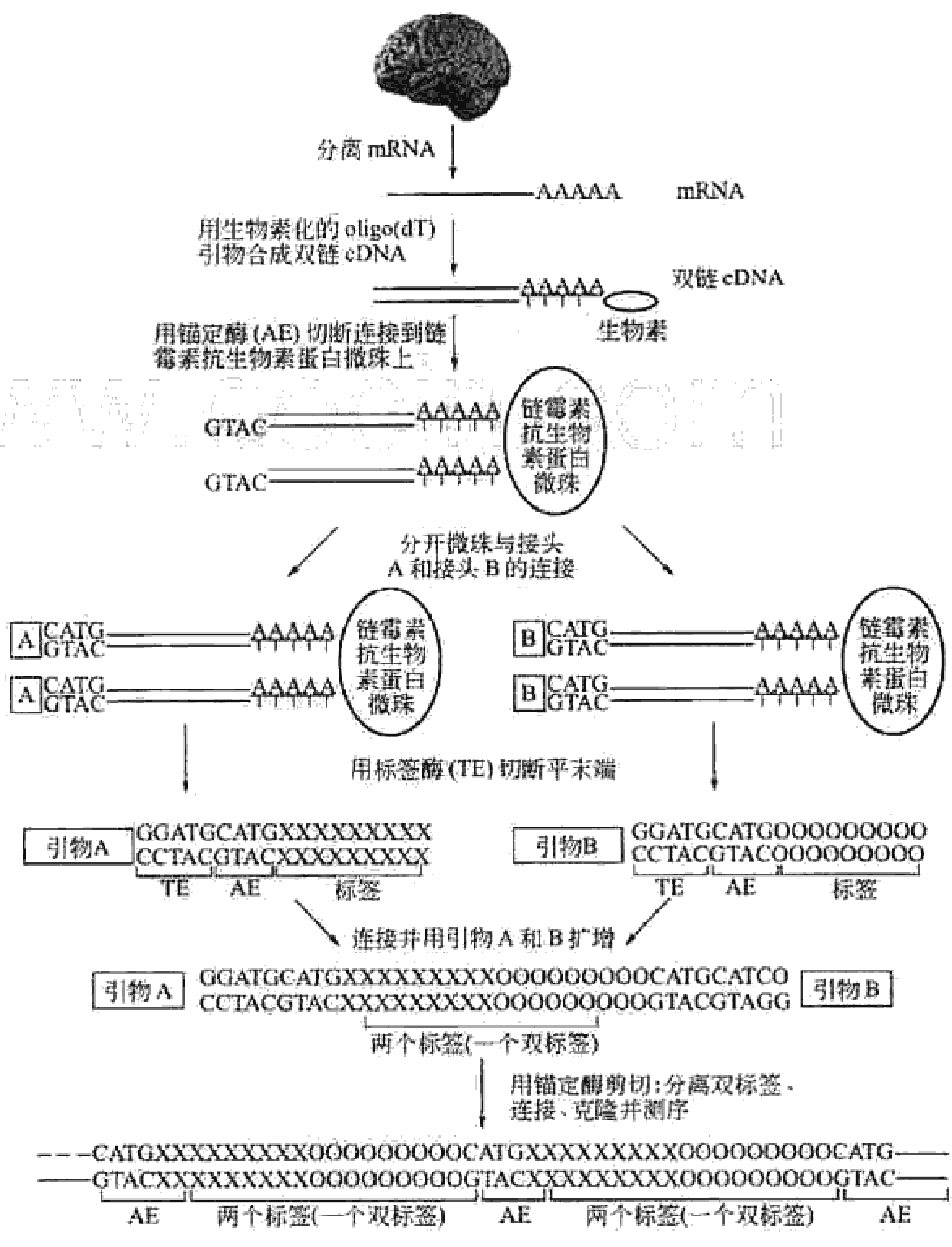


图6-9 基因表达序列分析（SAGE）的描述。先从样本（如脑）中分离出信使RNA，然后用生物素化的oligo(dT)引物合成双链cDNA，并用识别4个碱基的剪切细胞中大部分转录产物的限制性内切酶（“锚定酶”，AE）剪切。转录产物在3′端被固定在链霉素抗生物素蛋白的微珠（图中的大椭圆）上，并将含有限制性内切酶（“标签酶”，tagging enzyme，TE）识别位点的接头（A或B）添加上去。用标签酶截断绑定的克隆，从而释放出含有cDNA标签的接头（X记号在左边，O记号在右边）。然后将成对的标签连接起来，克隆并测序。每个标签代表一个转录产物中长9bp的片段。此图由Velculescu等（1995）修改得到，并经许可采用

现已构建出多种 SAGE 文库。文库中每个标签都可能对应于一个基因。假设标签位点的核苷酸分布是随机的，对于长度为 9bp 的标签，可被分辨的转录产物数量为 4^9 个，即 262144 个。实际上是通过 UniGene 将这些标签和基因对应起来。某些情况下，一个标签可能在不止一个基因中出现。其他情况下，一个基因则可能有不止一个标签位点（如对于可变剪切的转录产物会有多个标签对应于该基因）。SAGE 的一个假设是在 SAGE 文库中找到的标签数量是和该生物样本中的 mRNA 分子数成正比的。

SAGE 已被用于描述酵母转录组的性质（Velculescu 等，1997）。他们获得 4665 个基因的表达情况，其中大多数基因的功能尚未描述。数据表明大量转录产物很少表达，这和对 UniGene 簇的分析（表 6-1）是一致的。Zhang 等（1997）用 SAGE 分析了胃肠道肿瘤组织和相应正常组织中的基因表达。他们估计每个细胞类型中表达的不同转录产物数量在 14000～20000 个，一个细胞中各转录产物的表达水平在单拷贝到 5300 拷贝之间。对一个转录产物，将测量到的标签数除以总标签数，可作为对该基因表达丰度的估计。

在 NCBI 网站上可对 SAGE 文库进行电子检索。可以在已经建立了 SAGE 文库的组织之间进行基因表达的比较（Lash 等，2000；Lal 等，1999）。该网站上有来自 SAGE 文库的标签数据，还有将标签对应到基因后的注释数据（图 6-10）。对 SAGE 文库选择可以以类似于数字差别显示的方式进行（图 6-11）。对肺组织和脑组织的 SAGE 文库比较显示，表面活性剂（surfactant）、pronapsin A 和分泌珠蛋白（secretoglobin）在肺组织中对应的标签有数百个，而在脑组织中却没有（图 6-12）。脑组织中高表达的转录产物也被鉴别出来（图 6-12）。对表面活性剂的检查（点击其链接）显示这个特定标签（TGCCAGGTCT）惟一对应于表面活性剂基因（UniGene Hs. 177852），肺组织文库中对应于表面活性剂的该标签出现 50 次（图 6-13）。

网页中"Retrieve by sequence"（由序列获取）选项将完成传统 RNA 印迹的电子版本，即对某基因在多种组织中的表达水平进行评估。输入对应于 *RBP* 基因的 DNA 序列的编号后（图 6-14），相应的 SAGE 标签将被显示出来（图中的箭头 1）。点击该标签的链接后，显示出"retrieve by SAGE tag"网页（图 6-15）。包含该标签的文库，连同该标签出现的频率（以每 100 万个标签中出现的次数为单位）一起被列出来[图 6-15(a)]。这里，*RBP4* 在正常肝组织文库中出现最多。该页面还列出了含有这个标签的基因 [图 6-15(b)]。在这个例子中，该标签对应于 4 个不同的 UniGene 簇，但根据 NCBI 的标准（如相应基因组 DNA 数据是否已测出），只有 *RBP4* 簇被列为"reliable"（可靠）。

> USAGE 是一个基于网络的关系数据库，提供对 SAGE 数据的详细分析（http://www.cmbi.kun.nl/usage）（van Kampen 等，2000）。人类转录组图谱由 Amsterdam 大学开发（http://bioinfo.amc.uva.nl/HTM）。基于对 240 万 SAGE 标签的分析，Caron 等人（2001）描述了每个人类染色体的表达谱。他们在染色体范围内鉴别出数十个基因表达增长的区域。
>
> 另一个很有帮助的网站是 SAGE（http://www.sagenet.org）。除了提供数据、实验方案和 SAGE 图谱外，该站点还有详细的 SAGE 文献参考目录。

可以对 SAGE 文库进行电子比较的网络资源还有一些，如 USAGE（van Kampen 等，2000）和人类转录组图谱（Human Transcriptome Map）（Caron 等，2001）。

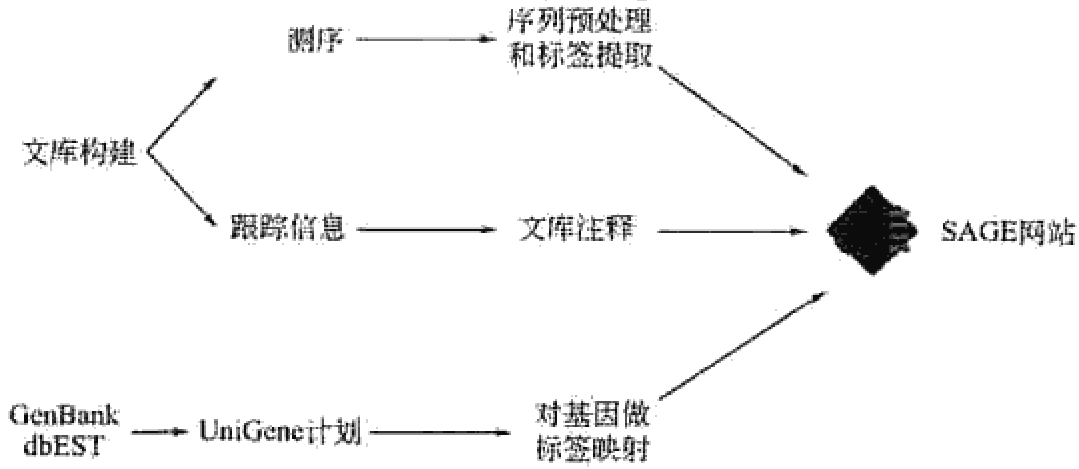


图 6-10 NCBI 的 SAGE 数据库的构建过程包含数据产生部分（上部：文库构建）和映射部分（下部）

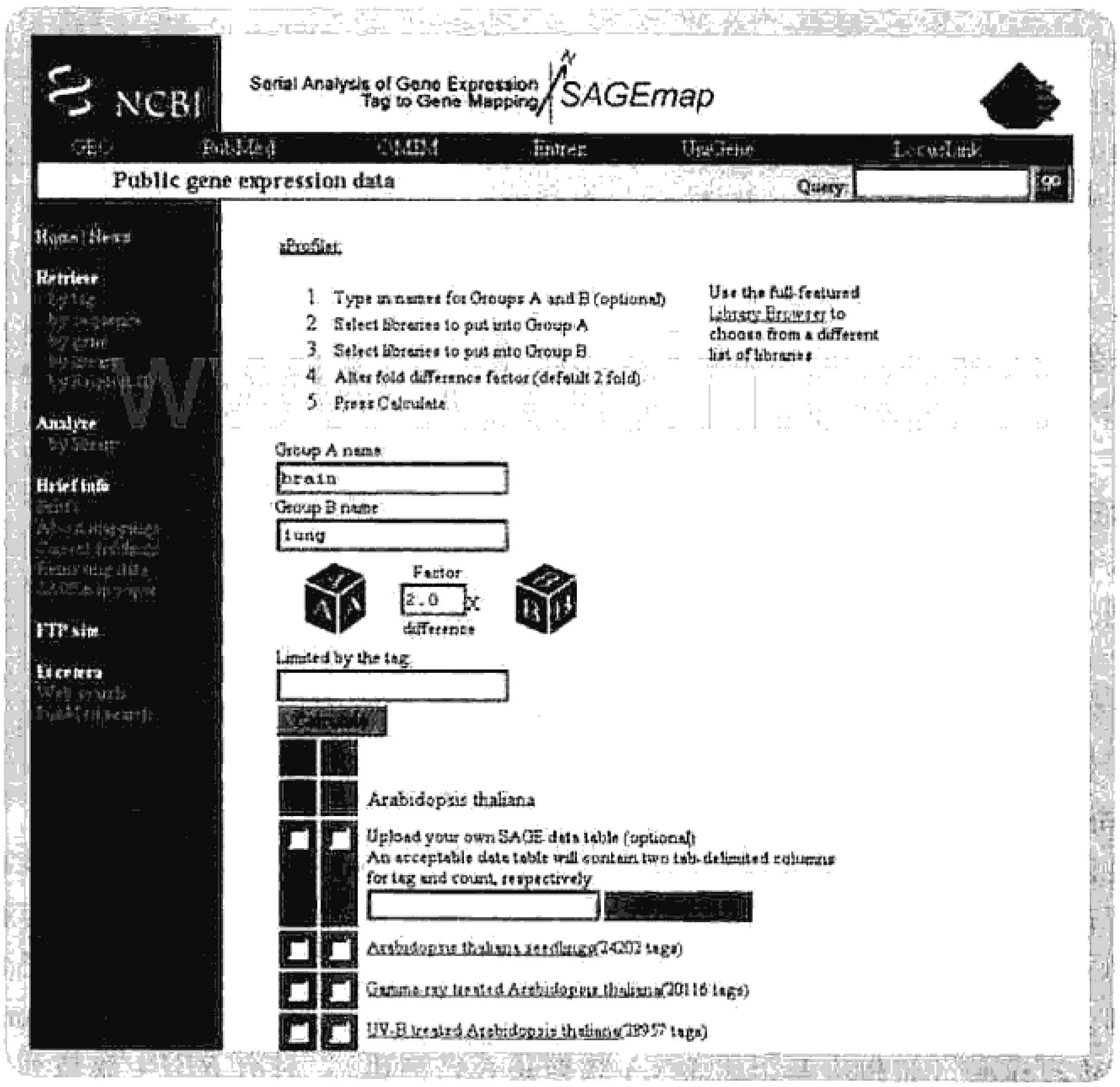


图 6-11 NCBI 的 SAGE 网站文库比较工具（从 http://www.ncbi.nlm.nih.gov/SAGE 点击“Analyze by library”进入）允许对两套 SAGE 的分析，从而比较基因表达谱。这里，标题 A 和 B 分别对应于人脑和肺的 SAGE 文库

Tag	Counts (brain) pool A	Counts (lung) pool B	p value	Unigene cluster	Cluster title
AACAGCAAAA	340	1	1.000	Hs.73133	MT3 metallothionein 3 (neurotrophic)
AACAGCAAAA	340	1	1.000	Hs.196550	LOC163590 AF464140
CGCAGCGGGT	0	270	1.000	Hs.322854	NAP1 (pronapsin A)
CAGGAGGAGT	63	1	1.000	Hs.13751	GRP58 glucose regulated protein
CAGGAGGAGT	63	1	1.000	Hs.40411	ESTs, Highly similar to S55507
CAACTAATTC	970	38	1.000	Hs.75106	CLU clusterin
CAACTAATTC	970	38	1.000	Hs.69997	ZNF238 zinc finger protein 238
CTTTGAGTCC	0	228	1.000	Hs.2240	SCGB1A1 secretoglobin (uteroglobin)
TCCCTATAAG	211	0	1.000		No reliable Unigene cluster found
GGCAAGAAAA	0	25	1.000	Hs.405528	RPL27 ribosomal protein L27
TGCCAGGTCT	0	50	1.000	Hs.177582	SFTPA2 surfactant, pulmonary-assoc.
GCGACCGTCA	447	8	1.000	Hs.273415	ALDOA (aldolase A, fructose-bispho.)
AAGAGCGCCG	87	0	1.000	Hs.75452	HSPA1A heat shock 70kDa protein 1A
AAGAGCGCCG	87	0	1.000	Hs.274402	HSPA1B heat shock 70kDa protein 1B
AAGCCAGCCC	87	0	1.000	Hs.1432	PRKCSH protein kinase C substrate
ACTCCCATAT	86	0	1.000	Hs.380027	CROC4 activator of c-fos promoter
TCCCATTAAG	253	1	1.000		No reliable Unigene cluster found
ACGAGGGGTG	85	0	1.000	Hs.6194	BCAN chondroitin sulfate proteoglycan
GTATGGGCCC	247	1	1.000	Hs.75184	CHI3L1 chitinase 3-like 1
GCCTGTCCCT	80	0	1.000	Hs.821	BGN (biglycan)

图 6-12 对脑和肺的 SAGE 文库的电子比较结果显示两个样本中表达丰度不同的转录产物对应的 SAGE 标签。本例中，B 组（肺）包括对应于表面活性剂和已知倾向于在肺组织表达的基因，其他标签对应于在脑组织中表达的基因，如神经营养金属巯蛋白基因

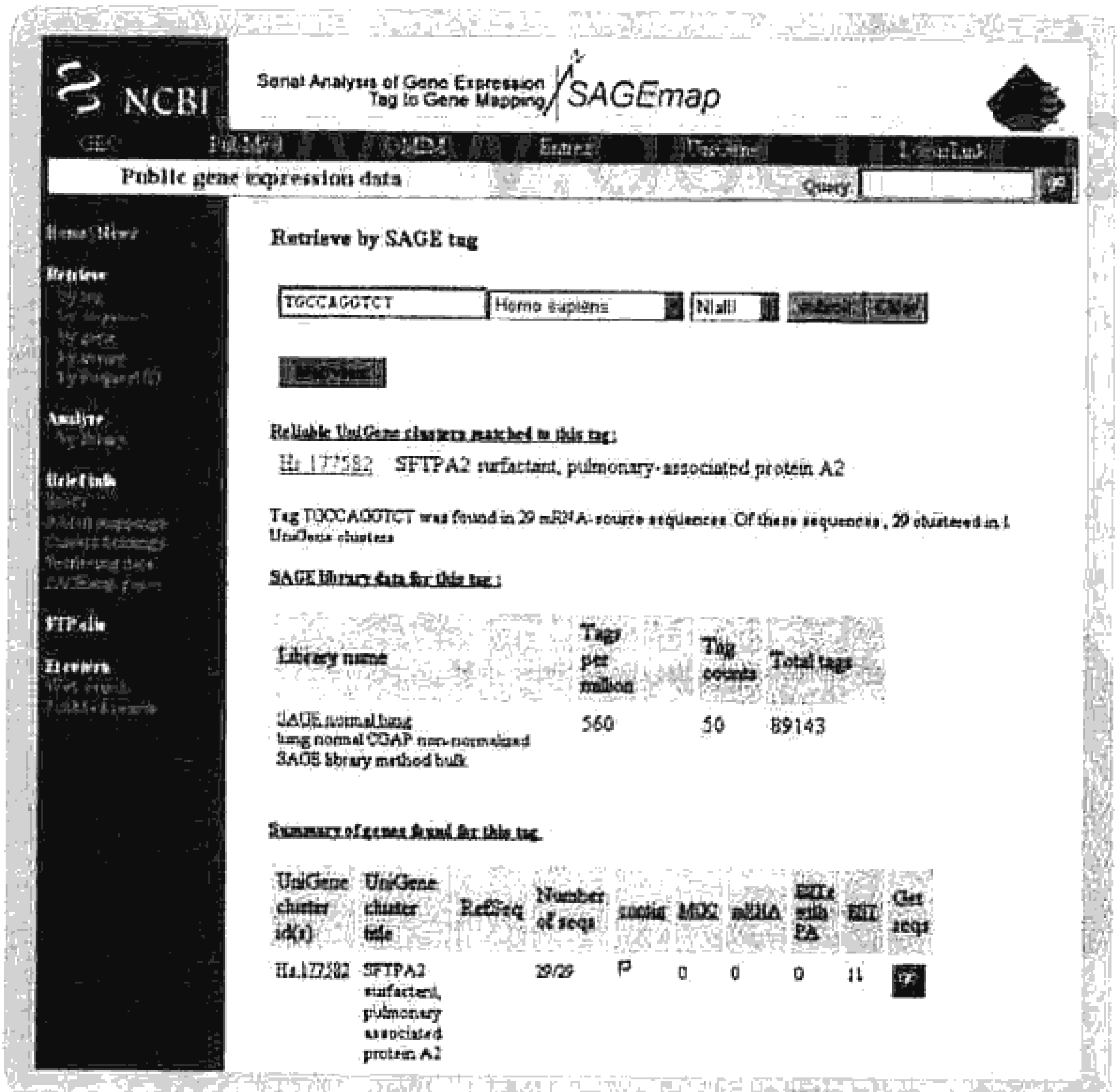


图 6-13 对应于表面活性剂的 SAGE 标签（见图 6-12）显示该转录产物已经被唯一地在正常肺组织中确定。可以通过在 NCBI 网站中点击“Retrieve by SAGE tag”获得这一结果

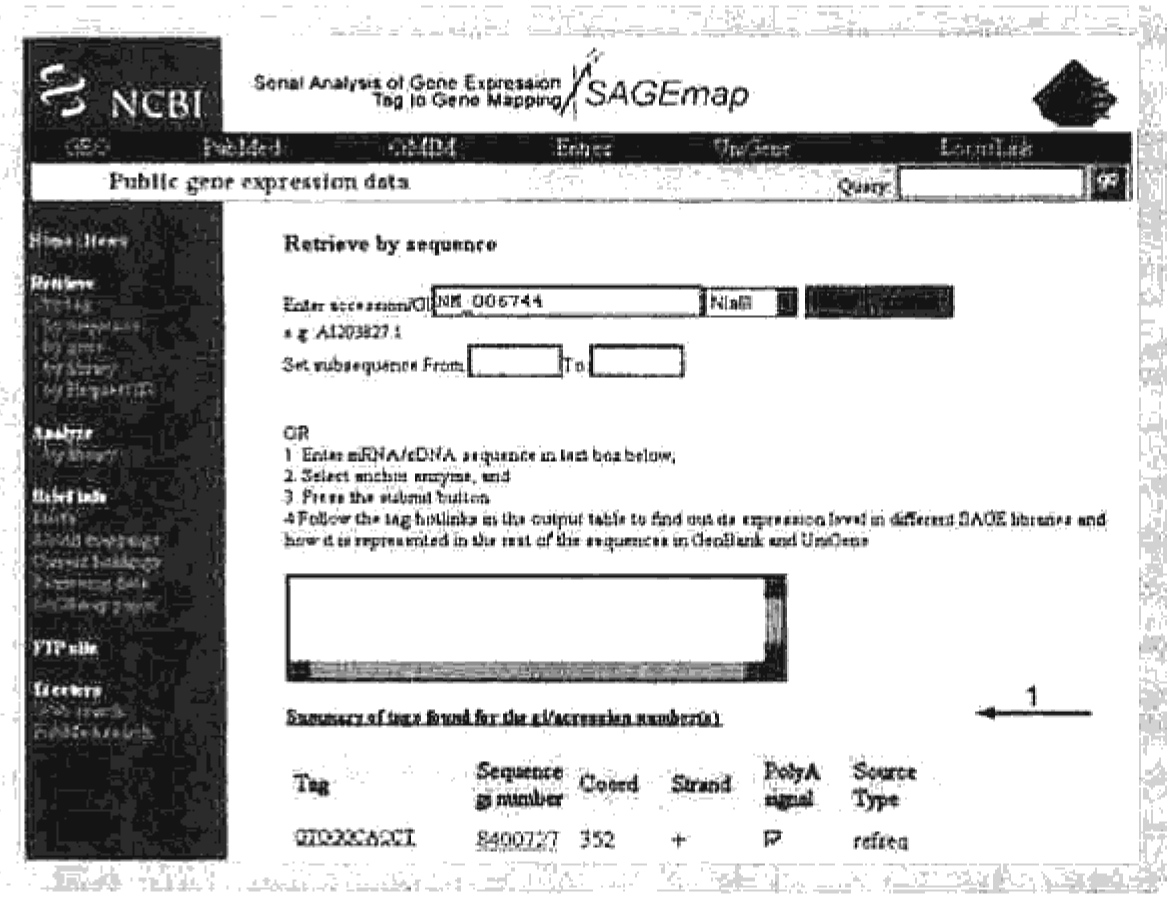


图 6-14 “Retrieve by sequence”工具使用户可以估计某转录产物在 SAGE 文库包含的组织中的分布和丰度。这里输入了 RBP4 的序列号（NM_006744），然后该标签的概况就显示出来（箭头 1）

(a) SAGE library data for this tag :

Library name	Tags per million		Tag counts	Total tags
SAGE Br N mammary gland epithelium ductal normal antibody purified CGAP non-normalized SAGE library method bulk	26		1	37642
SAGE normal pool(6th) brain normal SAGE CGAP non-normalized SAGE library method bulk	63		4	63208
SAGE Duke Kidney kidney normal CGAP non-normalized SAGE library method bulk	95		4	41857
SAGE Caco 2 adenocarcinoma colon SAGE CGAP non-normalized SAGE library method cell line	113		7	61667
SAGE normal liver liver normal CGAP non-normalized SAGE library method bulk	433		29	66861

(b) Summary of genes found for this tag

UniGene cluster id(s)	UniGene cluster title	RefSeq	Number of seqs	contig	MGC	mRNA	ESTs with PA	EST	Get seqs
Hs.418083	RBP4 retinol binding protein 4, plasma	NM_006744	3/6	☑	1	1	0	103	go
Hs.281462	FLJ14251 hypothetical protein FLJ14251		1/6						go
Hs.332566	CYP M cytochrome P450 monooxygenase		1/6						go
Hs.75313	AKR1B1 aldo keto reductase family 1, member B1 (aldose reductase)		1/6						go

图 6-15 搜索 *RBP4* 基因得到的虚拟 RNA 印迹实验的结果。这里显示的是标签（GTGGGCACCT）的数据，该标签对应于 *RBP4*。(a) 结果显示 *RBP4* 在 Caco-2（直肠癌）细胞系中出现 7 次，出现频率为每 100 万个中 113 个标签。如果一个文库中只鉴别出一个或几个标签（就像这里的一些情况），则要考虑是不是假阳性造成的。SAGE 标签可能被不恰当地分配给某个基因或文库。(b) 由这个标签找到的基因的概况显示 4 个 UniGene 簇含有这个序列

6.5 微阵列：基因表达的全基因组测量

DNA 微阵列（microarray）作为测量基因表达强有力的工具已经出现了。虽然 EST 测序计划和 SAGE 都可实现基因表达的高通量分析，但在对不同生物样本中 mRNA 丰度进行比较这一领域中，应用最广泛的还是微阵列。自从 Patrick Brown 和其同事在斯坦福大学，Jeffrey Trent 和其同事在 NIH 等做了开创性工作以来，微阵列的使用快速增长（De Risi 等，1996）。

微阵列由一块固体支持物（如显微镜载玻片或尼龙膜）和附着其上以规则的网格形式排列的序列已知的 DNA 组成。DNA 一般是 cDNA 或寡核苷酸，有时是其他材料（如基因组 DNA 克隆，见第 18 章）。被固定在阵列表面的 DNA 数量一般是几纳克。RNA 被从感兴趣的生物来源中抽提出来，如经过和未经过药物处理的细胞系，某物种的野生型和突变型样本，以及在不同时间点上被研究的同一样本。RNA（或 mRNA）常常被转换成 cDNA，做荧光或放射性标记后在阵列上杂交。杂交过程中，由来自生物样本中的 RNA 分子转换得到的 cDNA 将选择性地和微阵列表面上相应的核酸杂交。微阵列经洗涤后，经过图像分析和数据分析对探测到的信号进行量化。通过这个过程，微阵列技术实现了对阵列上数以千计的基因表达水平的同时测量。

对比于对单个蛋白或核酸分子的研究，“功能基因组学”这一术语是指全基因组范围内基因功能的大尺度分析。基于微阵列的基因表达实验构成了功能基因组学的核心。因具有产生大量信息的能力（表 6-4），科学界对微阵列积极性很高。值得注意的是，GenBank 中分子序列数据的迅速积累使我们可以获得数以千计功能未知的克隆。来自已知的基因和缺乏相关信息的 EST 的 DNA 序列都被置于微阵列上，以期探测其功能。这项技术潜力巨大，并且已经从中获得重大的生物学成果（见第 7 章）。

表 6-4 微阵列实验的主要优势

优　势	说　明
快速	可以在一周内获得 10000 个以上基因的表达水平数据
详细	整个酵母基因组可以呈现在一个芯片上
灵活	对应于任何基因的 cDNA 或寡核苷酸都可以呈现在芯片上

对微阵列技术局限性的认识也是很重要的（表 6-5）。微阵列实验的费用使很多研究者只做相对较少的对照实验以评估其结果的可靠性和正确性。其中存在很多质量控制因素，如很多研究者无法估计固定在阵列上的 DNA 的性质。有报道（anecdotal report）指出 DNA 序列中很大一部分是有问题的（如 DNA 被错误的注释，或者多个克隆被点在阵列上的同一个位点）。随着微阵列技术的成熟，出错率有望不断减小。

表 6-5 微阵列实验的主要劣势

劣　势	说　明
费用	对很多研究人员来说，其高昂的价格不容许作足够的重复和对照
RNA 的意义不清楚	基因表达的最终产物是蛋白质而不是 RNA
质量控制的不确定性	对大部分研究人员来说，不可能评估固着在微阵列上的 DNA 的性质，并且有很多图像分析和数据分析造成的人为影响

基因表达的最终产物不是 mRNA 而是蛋白质。细胞中 mRNA 表达水平和蛋白质水平的关

系还很不清楚。如果一个 mRNA 转录产物增加 2 倍，相应的蛋白质数量是否也增加 2 倍呢？对于细胞来说，基因表达变化的意义是什么？对这些问题的解答都还刚刚开始。目前，高通量的蛋白质分析与转录谱的研究相比，在技术上更难实施（尤其是蛋白质阵列，见第 8 章）。

有几个研究组报道在酿酒酵母和其他系统中，存在 mRNA 水平和相应蛋白质水平之间的弱的正相关（Futcher 等，1999；Greenbaum 等，2002）。Greenbaum 等（2002）对基因表达和蛋白质丰度数据集进行整合分析（meta-analysis），认为两者的表达水平之间存在广泛的一致性。

图 6-16 纵览了微阵列实验的各个步骤，这些步骤可以分为八个阶段。下面来看看每一步的具体情况。

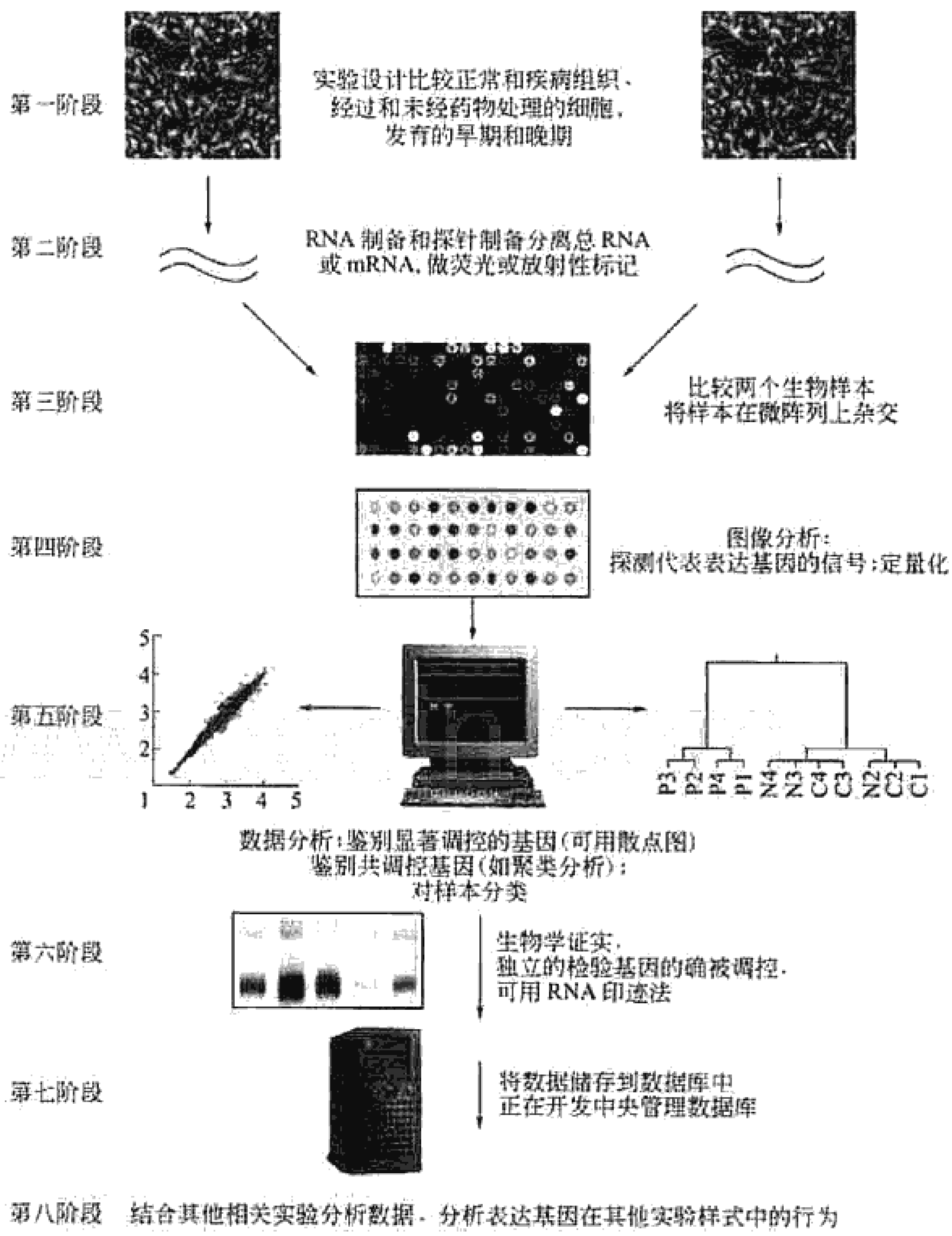


图 6-16 用微阵列产生高通量基因表达数据的过程纵览。第一阶段，选择需要对基因表达进行比较的生物样本。第二阶段，分离 RNA 并标记，一般使用荧光染料。这些样本被杂交到微阵列上，微阵列是含有互补 DNA 或对应于已知基因或 EST 的寡核苷酸的固体支持物。第四阶段，通过图像分析得到信号强度。第五阶段，分析表达数据，鉴别出被差异表达的基因（如使用散点图；第五阶段，图左），或基因及样本的聚类情况（第五阶段，图右）。在这些基于微阵列的发现的基础上，再进行独立的实验验证（第六阶段）。最后将微阵列数据储存到数据库中（第七阶段）以备大尺度分析（第八阶段）

6.5.1 第一阶段：微阵列的实验设计

在第一阶段，将总 RNA 或 mRNA 从样本中分离出来。值得注意的是，已经用作实验对象的物种范围十分广泛，如病毒、细菌、真菌和人类（见第 13～第 17 章）。初始材料的标准数量为数百毫克（湿重）或数烧瓶（flask）的细胞。对许多目前使用的微阵列，需要 1～25μg 总 RNA。如果考虑对 RNA 或 cDNA 产物进行扩增，初始材料的数量可以少很多。但是扩增后的产物不一定忠实反映初始 RNA 群体的情况。

微阵列实验的实验设计可以分成3部分（Churchill，2002），每部分都可以有不同因素的变动。

实验设计将在第7章作进一步讨论。

① 选出生物样本以供比较，如经过和未经药物处理的细胞系。如果使用的生物样本不只一个，则称为“生物重复”（biological replicates）。在选择实验对象进行处理时，合适的做法是对其随机分组。

Affymetrix公司的微阵列中，RNA被转换成cDNA，然后转录并生成生物素标记的cRNA。

② 将RNA提取出来并作放射性或荧光标记（一般以互补DNA为对象）。如果从同一个生物样本中得到两份RNA抽提物并杂交到微阵列上，则称为“技术重复”（technical replicates）。对于单个样本杂交到一个微阵列上的工艺，转录产物被放射性或一种荧光染料标记。图6-17(a)显示了一个实验设计范例，5个样本被杂交到微阵列上，其中4个样本可代表测定（assay）的不同时间点、不同的药理学处理，或任何其他模式。还可以使用一个经同一处理条件的多个样本混合而成的参考样本。

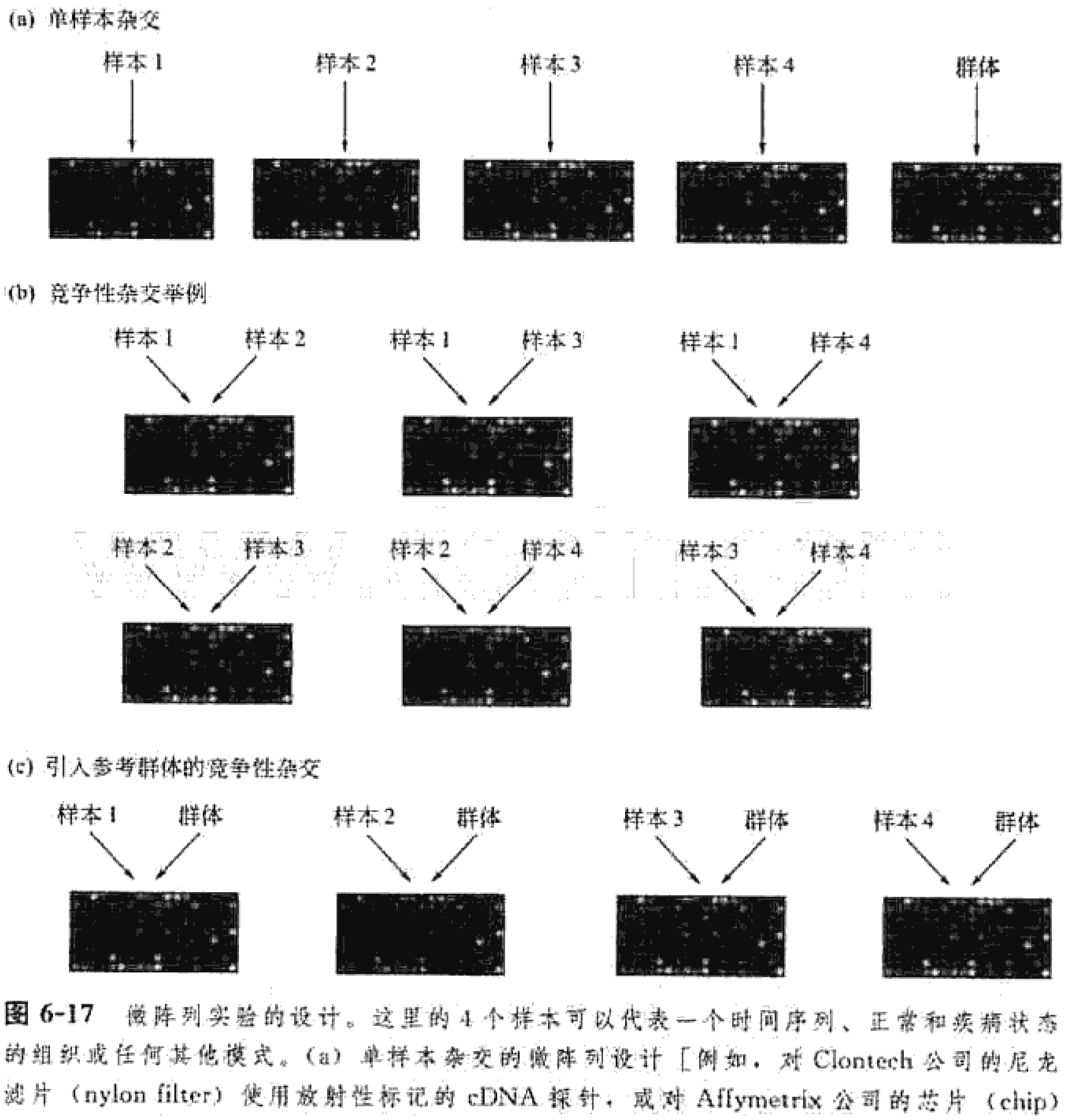


图6-17 微阵列实验的设计。这里的4个样本可以代表一个时间序列、正常和疾病状态的组织或任何其他模式。(a) 单样本杂交的微阵列设计［例如，对Clontech公司的尼龙滤片（nylon filter）使用放射性标记的cDNA探针，或对Affymetrix公司的芯片（chip）使用荧光染料标记cRNA探针］。每个样本各自杂交到一个芯片上，没有必要使用一个含有所有样本的群体（pool）。可以进行额外的杂交以增加实验重复次数。(b) 对于运用竞争性杂交的工艺，如NEN Life Science阵列，样本分别用Cy5（红色）和Cy3（绿色）染料标记并竞争性杂交。一套杂交实验可以包含样本的所有组合。不同芯片的亮度数据之间（如样本1和样本2）是不能进行比较的。(c) 为了对每两个样本之间进行比较分析，可以将每个样本和参考群体杂交

微阵列数据标准化和规范化（standardization and normalization of microarray data，SNOMAD）是一个免费的基于网络的微阵列数据分析方法集（Colantuoni 等，2002b），其中包括基于局部回归分析（local regression analysis）对位置引起的不准确的结果的修正。该程序计算了微阵列局部基础上（local basis）的本底。微阵列规范化的讨论见第7章。

对于应用竞争性杂交的微阵列工艺，实验设计有所不同[图 6-17(b) 和图 6-17(c)]。样本分别用 Cy5（一种红色染料）或 Cy3（绿色染料）标记。每个标记后的分子将和微阵列表面上的互补分子结合。如果转录产物在两个样本中的表达水平相近，则相应点的颜色为中间色（如黄色）。这个方法得到的是基因表达的相对比例，而不是绝对值。为了进行全面的数据分析，实验设计可囊括所有样本之间的两两比较［图 6-17(b)］。Churchill（2002）和其他一些人认为实验设计中参考样本的引入没有必要。

对于所有的复色实验模式，染料的互换是一种重要的对照。对两个生物样本，分别进行两次杂交，其中样本被染色两次，第一次用 Cy5，第二次用 Cy3，然后用于两次独立的杂交实验中。交换染色有助于消除由染料标记效率的不同而造成的人为差异。

③ 微阵列实验设计的第三方面是玻片上阵列元素的排列。理想状态下，阵列元素在玻片上随机排列。某些情况下，阵列元素被点样2次（见图6-19）。由于阵列元素的排列方式的影响或微阵列表面未被均匀清洗（或干燥），可能造成人为差异。

6.5.2 第二阶段：RNA 的制备与探针制备

可以用诸如 TRIzol（Invitrogen 产品）的这类试剂较容易地将 RNA 从细胞或组织中提纯出来。对某些微阵列的使用，需要将 RNA 进一步纯化为 mRNA [poly(A)$^+$ RNA]。如果要比较两个样本（如经过和未经药物处理的细胞），纯化 RNA 的过程中保持相似的条件十分关键。例如，对于细胞的培养，诸如在培养物中的天数和聚集百分比（percent confluence）等必须加以控制。

RNA 的纯度和质量应通过分光光度计（测量 A_{260}/A_{280} 值）和凝胶电泳加以评估。可用荧光染料［如 RiboGreen（Molecular Probes 产品）］定量测量产物。RNA 的纯度还可以通过 RNA 印迹或 PCR 检验。RNA 制备产物含有杂质（如基因组 DNA、rRNA、线粒体 DNA、糖类或其他大分子）可能是产生导致高本底的不纯探针或其他人为影响的原因。

RNA 被放射性或荧光染料标记，从而可对其进行探测。

6.5.3 第三阶段：将标记后的样本和 DNA 微阵列杂交

微阵列由尼龙膜或载玻片等固体支持物和固定在上面已知其序列的 DNA 片段组成。某些情况下，固着的 DNA 由约 5μg 排成行列的 cDNA（长度 100～2000bp）组成。其他情况下使用的是寡核苷酸（Lipshutz 等，1999）而非 cDNA，这已由 Affymetrix 公司用改进后的照相平版工艺（photolithograph）（Fodor 等，1991）实现。图 6-18 显示了寡核苷酸在薄片表面的分布。随用来固定 DNA 的固体支持物的性质不同，微阵列常被称为印迹（blot）、膜（membrane）、芯片（chip）或玻片（slide）。微阵列上的 DNA 被称为“目标 DNA”。在一个典型的微阵列实验中，对两个样本中的基因表达模式进行比较。每个样本中的 RNA 被

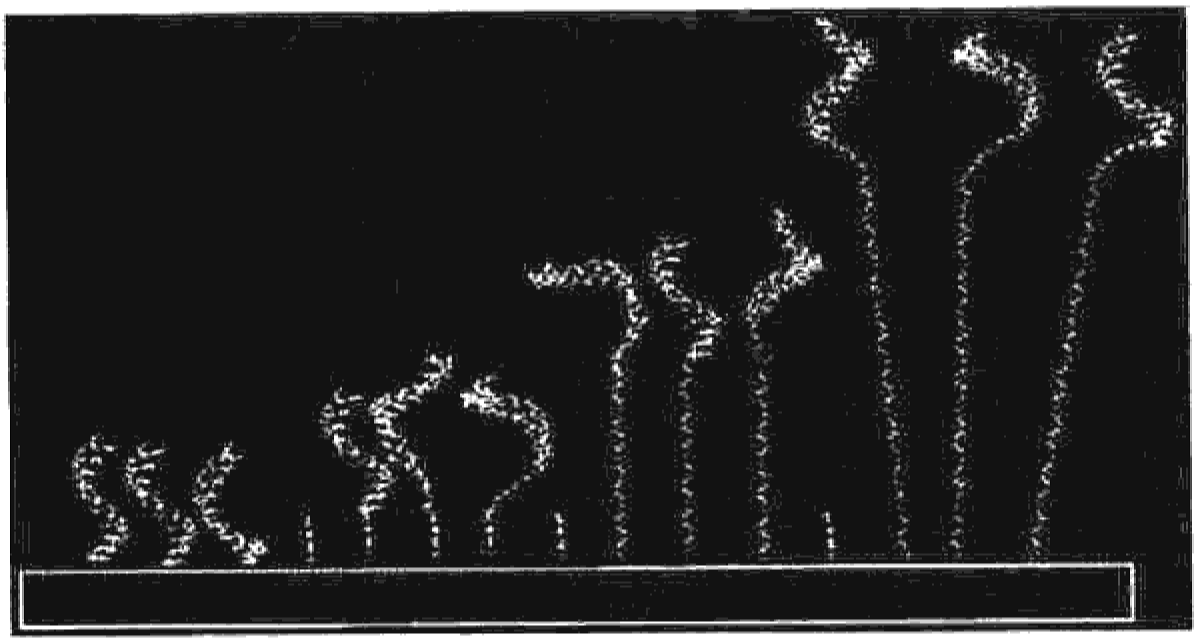


图 6-18　典型的微阵列芯片表面寡核苷酸的密度在载玻片上为 0.1pmol/mm^2，相当于每 3.9nm（39Å）一个分子（取自 Southern 等，1999）。Affymetrix 公司一个典型的微阵列对每个基因有 20 个长度为 25bp 的寡核苷酸，另外 20 个长度为 25bp 的寡核苷酸则加入了不匹配的碱基，预计无法和来自生物样本的 cRNA 杂交。通过分析对所有 40 个寡核苷酸的荧光信号分析，可以得知相应内源性转录产物的表达情况。其他微阵列平台，如 Agilent 或 Clontech 实验室的微阵列，使用的寡核苷酸最大长度为 80bp。经许可采用

施以荧光或放射性标记，形成“探针”。

将 RNA 转换成 cDNA 或 cRNA 并作荧光标记或放射性标记后，必须检验探针标记的效率，方法是使探针在滤片（filter）或玻片上的杂交持续一整夜，然后清洗。下一步是图像分析。

对用放射性和荧光标记的探针进行的微阵列实验特点进行比较是有益的。放射性探针被认为更灵敏，因为信号可以通过延长磷屏成像的时间而加强。放射性 cDNA 的合成需要的总 RNA 量一般是 1～25μg。在典型情况下，滤片先和一个样本杂交，剥离处理之后，再和另一个样本杂交。相反的是，荧光标记被认为对低表达和高表达的转录产物具有更宽的动态范围。荧光探针的合成有时需要多得多的 RNA（比如 100μg 总 RNA 或 1～5μg mRNA），而有些实验方案仅需要数微克的总 RNA。

照相平版工艺应用广泛，包括在微电子工业中用来将一些物质固定在固体表面。在微阵列技术中，通过在标准的寡核苷酸合成方法中使用对光不稳的（photolabile）核苷酸，可以在硅片表面原位合成数以千计的寡核苷酸。

很多研究人员将微阵列上的 DNA 称为探针，而称来自生物样本被标记的 DNA 为目标（target），这就产生了对探针和目标相反的定义，目前科学界对此尚未达成共识。在此一律将来自 RNA 或 mRNA 的标记过的材料称为“探针”。

6.5.4　第四阶段：图像分析

清洗微阵列之后，再对其进行图像分析以获得样本中各 mRNA 表达情况的定量描述（Duggan 等，1999）。在使用放射性探针（如使用典型的 [^{33}P] 或 [^{32}P] 同位素）的实验中，通过磷屏成像对结果定量化 [图 6-19，图 6-20(a)]。出售放射性标记的微阵列的公司也提供用于图像分析的软件，包括将各像素对准网格并人工调整网格使其与各点保持对齐。

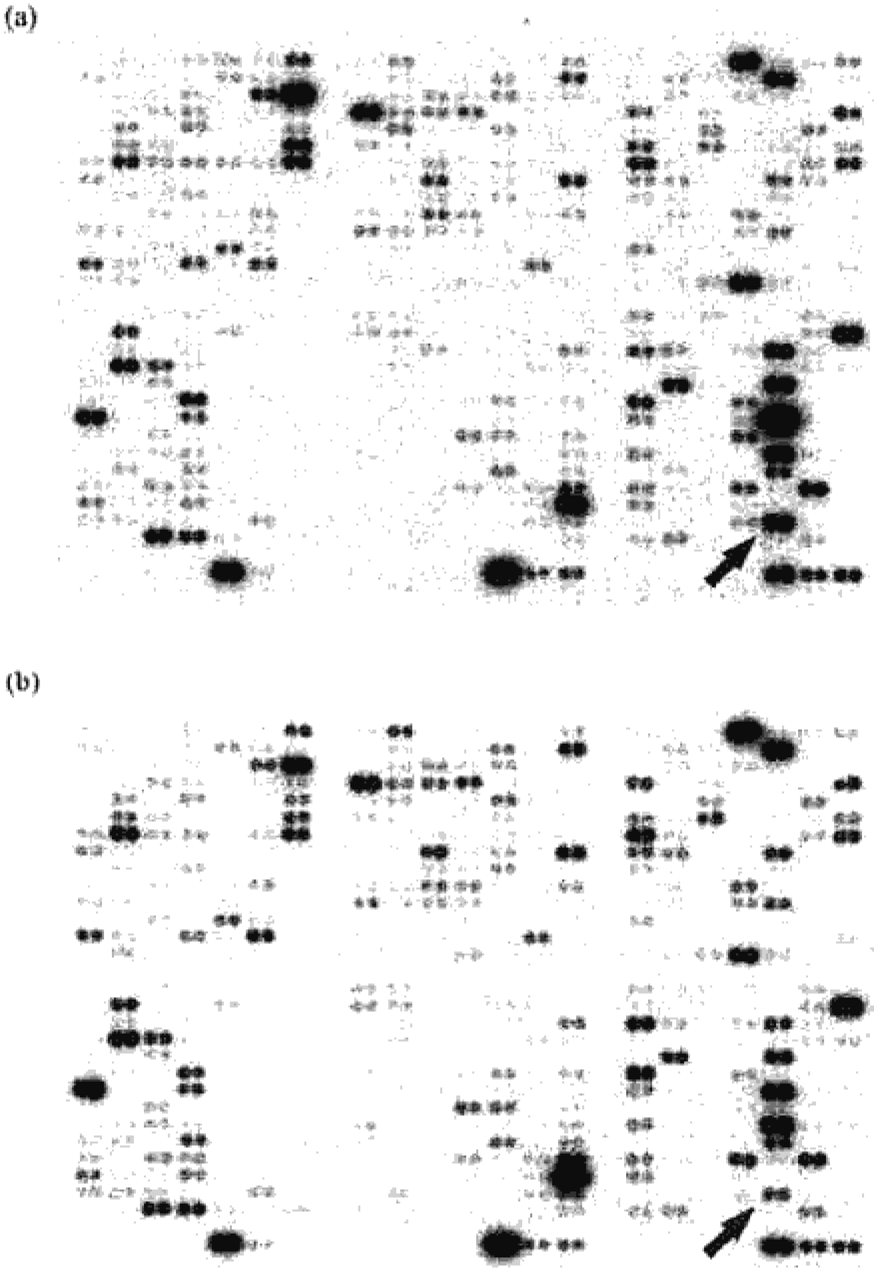


图 6-19 一个使用放射性探针的微阵列实验范例，每个阵列上有588个基因，每个基因为相邻的一对点。黑点代表表达水平高的基因。对滤片进杂交、清洗，然后在磷屏成像仪下暴露6h，输出结果中对应每个点的信号有一个数值。(a) 使用Clontech Atlas Neurobiology公司的微阵列，探针为取自患瑞特综合征（Rett syndrome）死亡后的小女孩的脑组织的cDNA。(b) 相应的对照组的表达谱。箭头所指的基因（β-晶状体球蛋白基因）在疾病组织中有上调，值得注意的是整体的基因表达谱在两种脑中是相似的

各点代表单个基因的表达水平。各点的强度被认为与样本中对应mRNA的量相关。然而其中可能有很多人为的影响。点的形状可能不规则。强度高的信号可能“溢出”到相邻的点使其信号被额外增强［图6-20(b)］。强度值接近本底的像素可能产生似是而非的高比值。例如，如果对照值高出本底100单位，实验值为200单位，则表明实验条件下表达上调2倍。然而，如果像素强度是50100对50200，则并不表明有调节发生。

基于荧光的实验中，阵列通过激光激发，然后测量荧光强度［图6-20(c)］。依次测得Cy5和Cy3通道的数据并得出基因表达水平的比值。

6.5.5 第五阶段：数据分析

对微阵列数据的分析可以鉴别出被差异调控的基因，也可以鉴别基因表达的显著模式。某些实验中成组的基因被共调控，显示出功能相关性。样本（而不是基因）被分析并分类。微阵列数据的分析将在第7章中介绍。

MIAME计划的介绍见微阵列基因表达数据库组织（microarray gene expression database group）网站（http://www.mged.org/）。

在微阵列数据分析标准化的尝试中，Alvis Brazma等（2001）来自17个科研机构的人员提议建立一个微阵列数据的储存和共享系统。微阵列实验最低限度信息计划（minimum information about a microarray experiment，MIAME）给研究人员提供了一个框架，记录了6个方面的信息：实验设计、微阵列设计、样本（及其制

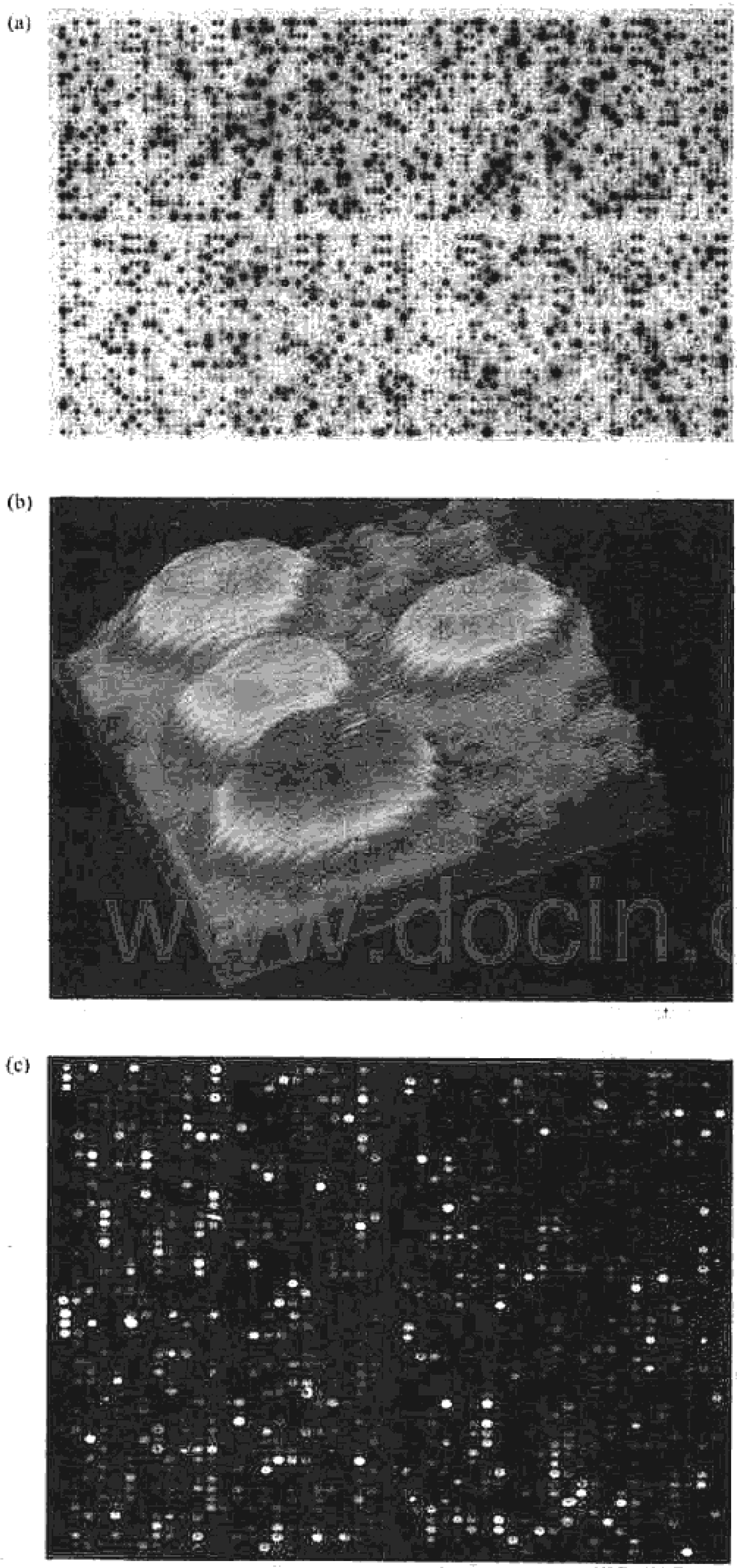


图 6-20 Research Genetics 公司的 GeneFilter 微阵列（探针为从唐氏综合征患者死后脑部海马区得到的 cDNA）。(a) 微阵列上共有 5000 种 cDNA，基因在所有阵列上的分布必须随机化。(b) 用 NIH 图像软件看到的 6 个信号点。图像分析软件必须详细说明每个信号的性质，包括一个强信号（图中左下角）“溢出”到弱信号（右下角）的可能性。(c) NEN Perkin-Elmer 公司的一种微阵列（MICROMAX，有 2400 个基因），用和图 6-19 中同一个瑞特综合征患者和相应对照组的脑组织样本做探针。这里使用的是经荧光标记和竞争性杂交的样本（见彩图）

备）、杂交过程、图像分析、为标准化做的调节。

6.5.6 第六阶段：生物学证实

微阵列实验的结果是数千个基因表达值的定量测量，数据分析往往显示出数十到数百个基因被显著调节，结果依赖于特定的实验模式和统计分析方法。如果对基因按调节水平排序，认为最前面那 1% 是“显著”调节的，则一个含有 10000 个单元的微阵列中将有 100 个显著调节的基因。也许细胞中这些基因的表达的确被调节了，也许这些基因之所以被鉴别出来只是人为因素或测量对象和测量过程中的随机波动的结果，所以至少对调节最显著的基因中的一部分进行独立的实验验证是很重要的。

6.5.7 第七阶段：微阵列数据库

许多科研人员一致认为微阵列数据一经发表就应该储存到公共数据库中。这样的资源刚开始出现（参见 Gardiner-Garden 和 Littlejohn 2001 年的综述）（表 6-6）。有人提议建立将微阵列数据和表达数据通过互补技术（complementary technologies）合并成基因表达数据库（如 Strachan 等，1997；Aach 等，2000）。初步的尝试主要来自欧洲生物信息研究所（European Bioinformatics Institure，EBI）和 NCBI。资料库的一个范例是 Stanford 微阵列数据库（Stanford Microarray Database），它提供了指向大量微阵列实验完整原始数据和处理后数据集的链接（图 6-21）。

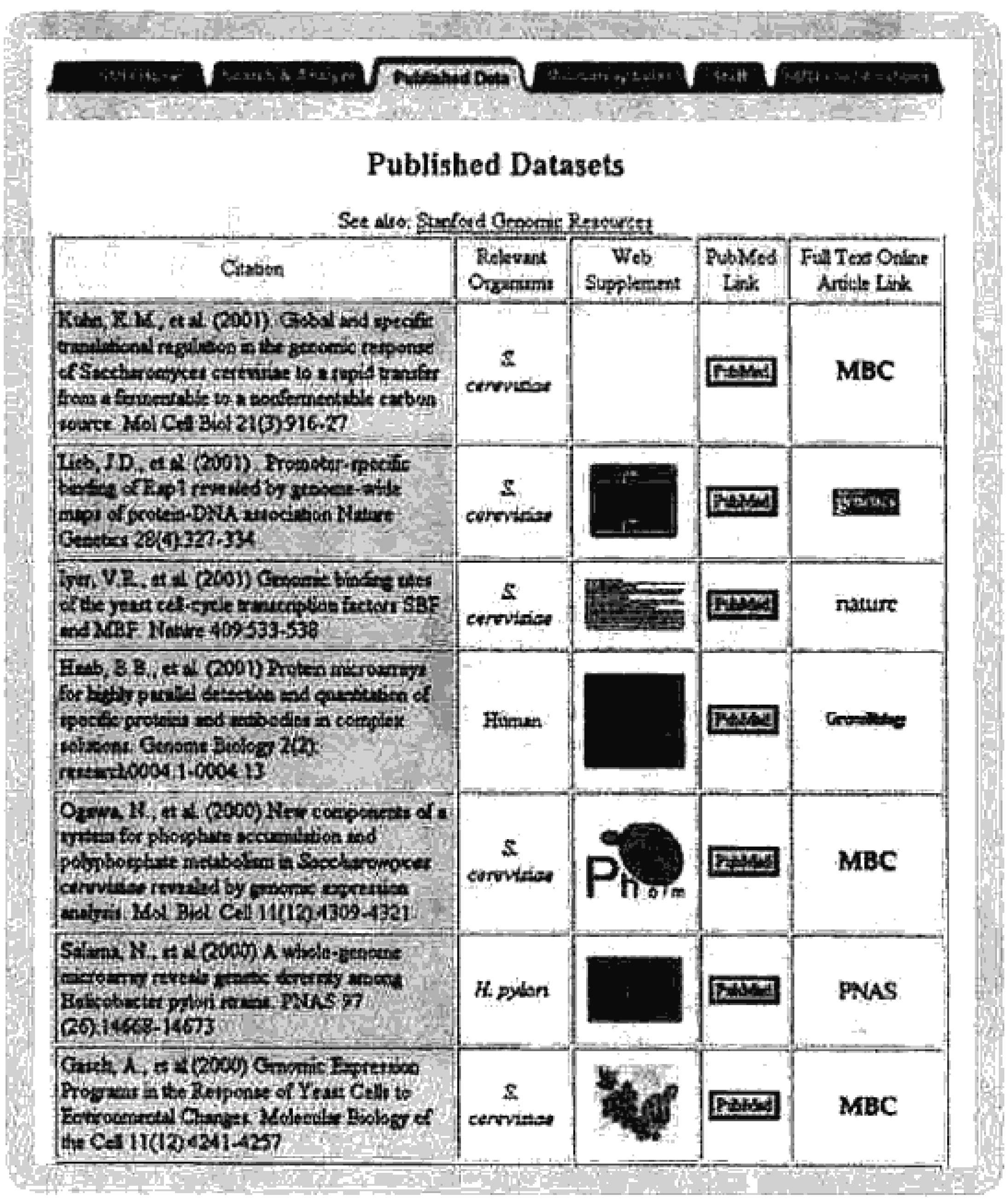


图 6-21 Stanford 微阵列数据库（http://www.dnachip.org/）提供的资源包含对已发表数据集的访问（如图所示），微阵列图像和数据分析软件

6.5.8 第八阶段：深入分析

可能所有的微阵列实验最终都会采纳统一的标准。这些研究中最大的不确定因素可能是各研究者分离的 RNA 质量和用于产生数据的微阵列性能。生物信息学领域当前的一个趋势是大量数据库的整合与相互参照，就像已经在分子序列数据库和蛋白质结构域数据库之间发生的情况那样。基因表达方面，由于可接受的标准的缺乏，可能限制从整合角度观察基因表达的程度。不过很有希望的是，将所有物种的每个基因都编入索引，其中除了分子序列和染色体位置等“稳定”数据，还包括基因对应的 mRNA 的“动态”信息。这些信息将包含各转录产物的丰度水平、基因表达的时间和区域位置以及不同状态下基因表达的行为等其他信息。第 7 章介绍了整合基因表达信息的一些初步尝试。

6.6 展望

所有生物中的基因都是在各种发育、环境或生理条件下表达的。功能基因组领域有赖于基因表达的高通量研究。这一新途径出现以前，典型研究仅限于单个基因在某一时刻的表达。功能基因组将揭示整个基因组的转录程序，从而实现对细胞功能的整体观察。

互补 DNA 微阵列和寡核苷酸微阵列出现于 20 世纪 90 年代中期，已经成为一个强有力并被广泛应用的工具，用于各种生物系统中基因表达的快速定量分析。随着阵列中出现的基因数量的增加、物种的增加和微阵列技术实验应用的不断发展，微阵列的使用将会更加广泛。

6.7 常见问题

对于应用诸如 EST 分析、SAGE 或微阵列这类技术的研究工作，有许多基本注意事项。mRNA 分子并没有被直接测量，而是先被转换成 cDNA，然后对 cDNA 进行序列分析，或经荧光或放射性标记物可视化后进行分析。被测量的物质的量是否对应于生物样本中 mRNA 的量，是需要考虑的一个重要问题。

① RNA（或 mRNA）被分离时，对于细胞中的 mRNA 分子整体是否有代表性？

② 如果是不同状态下的比较，RNA 是否在完全相同的条件下被分离？实验方案的不同可能导致人为的差别。

③ 任何样本中是否发生了 RNA 的降解？对于微阵列，相应地有其他因素。

④ 很多研究人员无法确认被固定在微阵列表面的分子。怎么知道被固定的是正确的克隆？又怎么知道阵列上各种 cDNA 或寡核苷酸的数量相当？有几项未公开的研究工作认为一些商品化微阵列上 10%～30%的基因是错误的。这就意味着被点样的克隆可能是错误的，或者某个克隆无意中被多次点到芯片上。

⑤ 微阵列表面测量到的每个点的可靠性如何？

面对关于微阵列的这一系列问题，一个反应是通过适当的实验设计才能得到可信的结果。在通过数据分析（见第 7 章）鉴别出被显著调节的基因后，独立的生化检验（如 RNA 印迹或 RT-PCT）对于结果的确认十分关键。

6.8 网络资源

网络资源见表6-6。

表6-6 微阵列数据集

数据集	说明	网址
AMAD	来源于斯坦福大学、加州大学伯克利分校和旧金山分校	http://www.microarrays.org/software.html
ArrayExpress	来源于EBI的Alvis Brazma等	http://www.ebi.ac.cn/arrayexpress/
ChipDB	来源于Whitehead研究所	http://young39.wi.mit.edu/chipdb_public/
ExpressDB	位于Harvard大学；是包含酵母RNA表达数据的关系数据库	http://arep.med.harvard.edu/ExpressDB/
Gene Director	来源于Biodiscovery公司	http://www.biodiscovery.com/
GeNet	来源于Silicon Genetics公司	http://www.sigenetics.com/
GeneX	来源于NCGR(National Council for Geocosmic Research)	http://genex.ncgr.org/
GEO	来源于NCBI的"Gene Expression Omnibus"	http:/www.ncbi.nlm.nih.gov/geo/
GXD	来源于Jackson实验室	http://www.informatics.jax.org/
MAdb	美国国家癌症研究院	http://madb.nci.nih.gov/
MaxdSQL	Manchester大学	http://www.bioinfo.man.ac.uk/microarray/maxd
RAD	Pennsylvania大学	http://www.cbil.upenn.edu/rad2/servlet
Stanford Microarray Database	Stanford大学	http://www.dnachip.org/

6.9 问题讨论

1. 若对一个人类细胞系测量其中药物诱导下基因表达的变化，分别用扣除法、SAGE和微阵列手段的有利因素和不利因素是什么？如果要研究的是像寄生物这样研究不多的生物的基因表达，答案有何不同？

2. 使用一个微阵列时，怎样评估被沉积在阵列表面的物质？怎样知道DNA的长度和组成与制造商所制定的一致？

6.10 习题

1. 执行以下数字差别显示：

① 进入UniGene网站（http://www.ncbi.nlm.nih.gov/UniGene)；

② 选择"Homo sapiens"；

③ 点击"library differential display"；

④ 点击某个（或某些）脑文库（brain library)，然后点击"accept changes"；

⑤ 选择另一组文库进行比较。

2. 执行数字SAGE：

① 进入http://www.ncbi.nlm.nih.gov/并点击"Serial Analysis of Gene Expression"；

② 点击“Analysis by library”；
③ 比较两组 SAGE 文库（如脑和卵巢）；
④ 进入 Entrez 或者 LocusLink 并选择一个 DNA 序列（任何一个均可）；
⑤ 将该 DNA 序列拷贝到剪贴板并返回 SAGE 网页；
⑥ 点击“Virtual Northern”；
⑦ 粘贴序列并提交。

6.11 自测题

1. RNA 处理阶段不包括以下哪个步骤？
(a) 剪切；
(b) 出核（export）；
(c) 甲基化；
(d) 监控（surveillace）

2. 数字差别显示被用来比较 UniGene 的 cDNA 文库中的表达序列标记（EST）。EST 还出现在微阵列中，对 EST 最好的描述是以下哪一项？
(a) 一簇非冗余序列（长约 500bp）；
(b) 一段在基因组中多次重复的 DNA 序列；
(c) 对应于通过对 cDNA 测序获得的表达基因的序列；
(d) cDNA 的一个“标签”（如一个 DNA 片段），对应于尚未被鉴别的转录产物

3. UniGene 中的簇大小范围从很小（如 1）到相当大（如超过 10000）。簇大小为 1 代表什么情况？
(a) 已鉴别出的一条序列，具有与其相关的大量 EST 转录产物（如超过 10000）；
(b) 已鉴别出的一条序列，对应于仅被表达 1 次的基因；
(c) 已鉴别出的一条序列（假定是一个 EST），和另一个已知序列相匹配（从而作为一个 UniGene 簇被鉴别）；
(d) 已鉴别出的一条序列（假定是一个 EST），被认为对应于一个已知基因，但在 UniGene 中没有任何匹配序列（也就是和任何 EST 都不匹配）

4. 在比较两个 cDNA 文库或 SAGE 文库时，可以鉴别出在其中之一存在差异表达的基因。Fisher's exact test 可被用来检验以下零假设：
(a) 任何一个基因在两个文库中（如心脏和肝脏）的序列数量相等；
(b) 任何一个基因在两个文库中（如心脏和肝脏）的序列数量不等；
(c) 两个文库中（如心脏和肝脏）的序列总数相等；
(d) 两个文库中（如心脏和肝脏）的序列总数不等

5. 分析 cDNA 文库时需要注意的是：
(a) 文库可能来源于不同组织；
(b) 文库可能包含数千个序列；
(c) 文库的归一化方式可能不同；
(d) 文库可能包含大量很少被表达的转录产物

6. 寡核苷酸微阵列相对于 cDNA 微阵列有何优势？
(a) 两个样本可以被同时竞争性地杂交到同一个微阵列上；
(b) 便于实验者鉴别微阵列上每个基因的身份；
(c) 有可能鉴别出可变剪切转录产物的表达；
(d) 灵敏度更高

7. 微阵列有一块固体支持物，被固定在上面的一般是
(a) DNA；
(b) RNA；
(c) 基因；

(d) 转录产物

8. 在微阵列研究中，RNA 样本通常由 cDNA 或 cRNA 转化而来，用来可视化的标记是

(a) 放射性或磷光；

(b) 放射性或荧光；

(c) 放射性或 RNA 探针；

(d) 放射性或 DNA 探针

9. MIAME 计划的目标是提供：

(a) 一个统一标准的系统，描述微阵列的制造；

(b) 一个统一标准的系统，描述微阵列实验，从设计到杂交乃至图像分析；

(c) 一个统一标准的系统，描述微阵列探针的制备，包括荧光和放射性标记；

(d) 一个统一标准的微阵列数据库系统，包括数据存储、分析和显示标准

10. 数以千计的基因的表达情况可以用 cDNA 文库、SAGE 和 DNA 微阵列测量。DNA 微阵列特有的一个优势是：

(a) 表达水平可以被量化；

(b) 可以在两种感兴趣的特定条件下测量数千个基因的表达水平；

(c) 用于比较两种特定条件下数千个基因的表达水平时，比其他两种方法更实用；

(d) 可被用来测量基因组中所有基因的表达水平

6.12 推荐读物

Mark Adams 和 J. Craig Venter 等（1991）关于 cDNA 研究早期的一个报告介绍了对从人脑 cDNA 中随机选择的 600 多个序列的分析。引人注目的是，其中 337 个是新基因。在进一步的筛选中他们鉴别出了 6000 多个人类基因（Adames 等，1992，1993）。NCBI 的 Gregory Schuler（1997）有一个对 EST 的评述。

本章描述了基因表达序列分析方法（SAGE），Johns Hopkins 大学的 Victor Velculescu 和同事（1995）引入了这项技术，图 6-5 即取自该文。Velculescu 和同事（1997）进一步用 SAGE 对酵母基因表达做出了富于洞察力的研究。在《科学》杂志 2001 年 2 月的人类基因组专刊上，Caron 等描述了基于 SAGE 数据的人类转录组图谱，显示出染色体位置造成的基因表达变化。

对于微阵列技术的介绍有很多（Duggan 等，1999；Hegde 等，2000；Zhang，1999）。Stephen Fodor 和同事（1991）描述了将寡核苷酸固定在固体支持物上的过程。Affymetrix 公司采用了这一方法，包括基于磷屏成像技术的光引导的、可空间定位的寡核苷酸大批量并行化学合成的方法。该文介绍了应用这项技术将肽链和寡核苷酸固定到固体表面的方法。

参 考 文 献

Aach, J., Rindone, W., and Church, G. M. Systematic management and analysis of yeast gene expression data. *Genome Res.* **10**, 431-445 (2000).

Adams, M. D., et al. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252**, 1651-1656 (1991).

Adams, M. D., et al. Sequence identification of 2,375 human brain genes. *Nature* **355**, 632-634 (1992).

Adams, M. D., Kerlavage, A. R., Fields, C., and Venter, J. C. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nat. Genet.* **4**, 256-267 (1993).

Bishop, J. O., Morton, J. G., Rosbash, M., and Richardson, M. Three abundance classes in Hela cell messenger RNA. *Nature* **250**, 199-204 (1974).

Bonaldo, M. F., Lennon, G., and Soares, M. B. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res.* **6**, 791-806 (1996).

Bouton, C. M. and Pevsner, J. Dragon: Database Referencing of Array Genes Online. *Bioinformatics* **16**, 1038-1039 (2000).

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Caus-

ton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., Vingron, M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).

Caron, H., et al. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science* **291**, 1289–1292 (2001).

Carulli, J. P., et al. High throughput analysis of differential gene expression. *J. Cell. Biochem. Suppl.* **31**, 286–296 (1998).

Celotto, A. M., and Graveley, B. R. Alternative splicing of the *Drosophila* Dscam pre-mRNA is both temporally and spatially regulated. *Genetics* **159**, 599–608 (2001).

Churchill, G. A. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics Suppl.* **32**, 490–495.

Claverie, J. M. Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.* **21**, 1821–1832 (1999).

Colantuoni C., Henry G., Zeger S., Pevsner J. SNOMAD (Standardization and NOrmalization of MicroArray Data): web-accessible gene expression data analysis. *Bioinformatics.* **18**, 1540–1541 (2002b).

Colantuoni C., Henry G., Zeger S., Pevsner J. Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. *Biotechniques.* **32**, 1316–1320 (2002a).

DeRisi, J., et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **14**, 457–460 (1996).

Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. Expression profiling using cDNA microarrays. *Nat. Genet.* **21**, 10–14 (1999).

Fodor, S. P., et al. Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767–773 (1991).

Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S., and Garrels, J. I. A sampling of the yeast proteome. *Mol. Cell. Biol.* **19**, 7357–7368 (1999).

Gardiner-Garden, M., and Littlejohn, T. G. A comparison of microarray databases. *Brief Bioinform.* **2**, 143–158 (2001).

Greenbaum, D., Jansen, R., and Gerstein, M. Analysis of mRNA expression and protein abundance data: An approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics* **18**, 585–596 (2002).

Hastie, N. D., and Bishop, J. O. The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* **9**, 761–774 (1976).

Hegde, P., et al. A concise guide to cDNA microarray analysis. *Biotechniques* **29**, 548–550, 552–554, 556 passim (2000).

Lal, A., et al. A public database for gene expression in human cancers. *Cancer Res.* **59**, 5403–5407 (1999).

Lash, A. E., et al. SAGEmap: A public gene expression resource. *Genome Res.* **10**, 1051–1060 (2000).

Lee, Y., et al. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res.* **12**, 493–502 (2002).

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. High density synthetic oligonucleotide arrays. *Nat. Genet.* **21**, 20–24 (1999).

Maniatis, T., and Reed, R. An extensive network of coupling among gene expression machines. *Nature* **416**, 499–506 (2002).

Maquat, L. E. Molecular biology. Skiing toward nonstop mRNA decay. *Science* **295**, 2221–2222 (2002).

Miller, O. L., Hamkalo, B. A., and Thomas, C. A. Visualization of bacterial genes in action. *Science* **169**, 392–395, 1970.

Milner, R. J., and Sutcliffe, J. G. Gene expression in rat brain. *Nucleic Acids Res.* **11**, 5497–5520 (1983).

Modrek, B., and Lee, C. A genomic view of alternative splicing. *Nat. Genet.* **30**, 13–19 (2002).

Quackenbush, J., et al. The TIGR Gene Indices: Analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* **29**, 159–164 (2001).

Sagerstrom, C. G., Sun, B. I., and Sive, H. L. Subtractive cloning: Past, present, and future. *Annu. Rev. Biochem.* **66**, 751–783 (1997).

Schmitt, A. O., et al. Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res.* **27**, 4251–4260 (1999).

Schmucker, D., et al. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**, 671–684 (2000).

Schuler, G. D. Pieces of the puzzle: Expressed sequence tags and the catalog of human genes. *J. Mol. Med.* **75**, 694–698 (1997).

Southern, E., Mir, K., and Shchepinov, M. Molecular interactions on microarrays. *Nat. Genet.* **21**, 5–9 (1999).

Stekel, D. J., Git, Y., and Falciani, F. The comparison of gene expression from multiple cDNA libraries. *Genome Res.* **10**, 2055–2061 (2000).

Strachan, T., Abitbol, M., Davidson, D., and Beckmann, J. S. A new dimension for the human genome project: Towards comprehensive expression maps. *Nat. Genet.* **16**, 126–132 (1997).

van Kampen, A. H., et al. USAGE: A web-based approach towards the analysis of SAGE data. Serial Analysis of Gene Expression. *Bioinformatics* **16**, 899–905 (2000).

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).

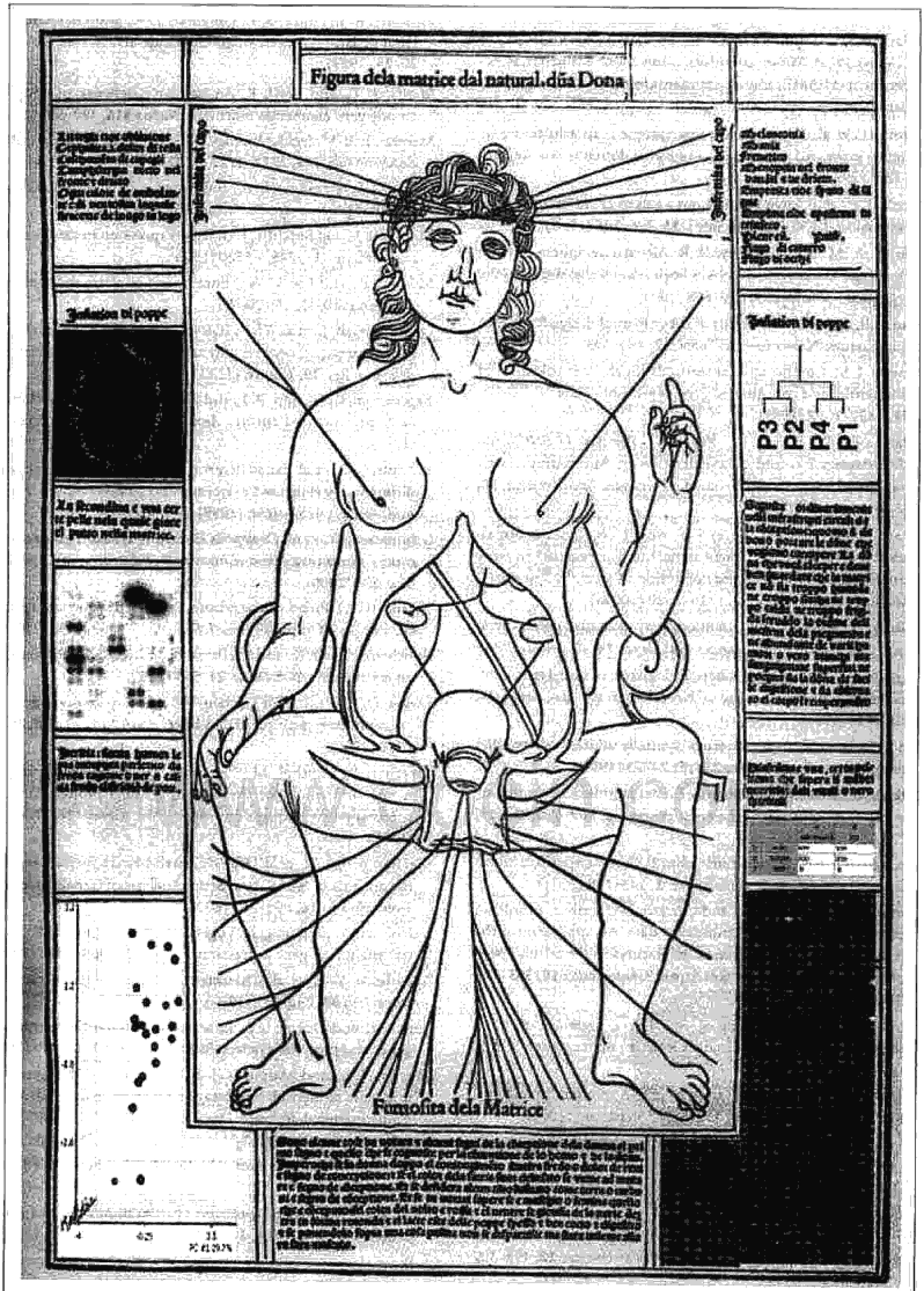
Velculescu, V. E., et al. Characterization of the yeast transcriptome. *Cell* **88**, 243–251 (1997).

Vietor, I., and Huber, L. A. In search of differentially expressed genes and proteins. *Biochim. Biophys. Acta* **1359**, 187–99 (1997).

Watson, J. B., and Margulies, J. E. Differential cDNA screening strategies to identify novel stage-specific proteins in the developing mammalian brain. *Dev. Neurosci.* **15**, 77–86 (1993).

Zhang, M. Q. Large-scale gene expression data analysis: A new challenge to computational biologists. *Genome Res.* **9**, 681–688 (1999).

Zhang, L., et al. Gene expression profiles in normal and cancer cells. *Science* **276**, 1268–1272 (1997).



基因芯片的雏形出现在 20 世纪 90 年代，首先是由斯坦福大学的 Patrick Brown 实验室然后是美国 NIH 研究开发的。基因芯片的含义可以由 Ketham 的书“Fasciculo di Medicina”中的孕妇像来形象地表示

第 7 章

基因表达：芯片数据分析

7.1 引言

随着科技的进步，DNA 芯片实验（DNA microarry experiments）已经成为现在最重要的一种分析基因表达的工具。当某个芯片实验完成，并且得到了相应的数据后，实验人员最想知道的问题是：在这个实验里，有哪些基因相对于它们自身的正常表达有非常明显的高或低的表达？这个问题的回答可以用推测统计学（inferential statistics）方法来解决，该方法是数据分析的一个分支，它主要是用概率论的方法来判定某个基因被调控重要性的可靠程度，该方法的主要步骤包括：

① 用电子数据表格列出所有在芯片上出现的基因和用来代表这些基因表达量高低的数据值；

② 用散点分析图方法（见下文）可以帮助我们快速绘出大部分调控基因的活动状况；

③ 用 t-检验可以来描述基因被调控的可能性。

另外需要知道的基本问题还有，在这个实验里，通过所有的基因表达数据可以得到基因表达的什么信息（或者模式、图谱等）。这种类型的问题可以用记述统计学（descriptive statistics）方法或者探测分析（exploratory analysis）的方法来解决。聚类树方法（clustering trees）可以表现不同样本之间（例如正常细胞 vs. 染病细胞），以及基因之间或者不同样本和基因之间的关系。其他分析基因表达的工具还包括主成分分析（principal components）、多维排列（multidimensional scaling）、自组织图（self-organizing maps）等方法。我们在以后的讲解中将利用所有的这些方法来分析芯片数据。

芯片实验需要解决的基本问题是测量一个生物样本成千上万的基因的表达水平。在这样的芯片实验中，经常需要技术性的重复（technical replicates）（例如，检测同种材料基因表达量的高低需在独立的多个芯片试样中重复），这也是通常基因芯片实验花费成本较高的原因。另外，在技术性重复的同时，也存在生物学重复（biological replicates）（例如，检测多个细胞系的基因表达水平，每个细胞系都会有一个实验处理或对照处理），这样的重复主要就是针对芯片中的大批量基因。对生物学家而言，最大的挑战就是要能够利用相应的统计学

技术来决定在基因表达数据中哪些改变是相关的。在这种对芯片数据的分析中，没有哪个方法是最好的，因为随着研究的深入，对芯片数据分析的工具也越来越多，分析结果也会越来越精确。

在探讨芯片实验的过程中，首先我们将不考虑芯片分析平台的问题，而是了解怎样分析芯片实验得到的数据（图 7-1），该矩阵的第二列含有的字母代表基因的名称，第二行

		1	2	3
		cell.line.C1	C2	C3
1	actin	100	150	105
2	tubulin	200	200	200
3	RBP	0	0	50

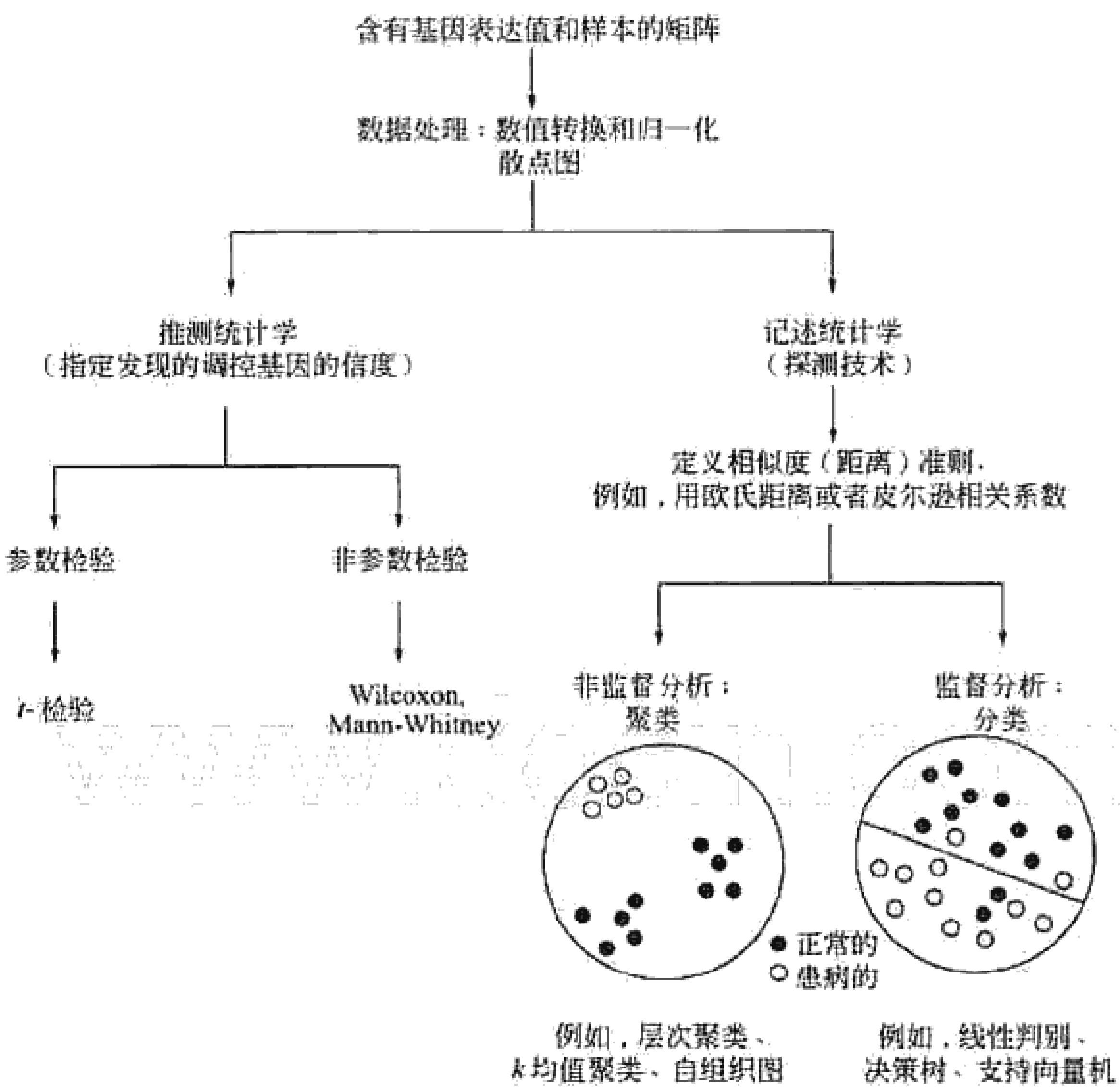


图 7-1 芯片表达数据分析概要，首先创建含有基因名称和样本名称的数据矩阵，然后对数据进行预处理（例如，芯片的总亮度归一化或把为负亮度值的数据点去除）。两种类型的统计学方法的应用：推测统计学方法通过假设检验对数据进行总结，例如，某一研究的基因受调控表达量相对正常表达量著提高［判断的标准是概率的高低，P 值（P value）越高说明其可靠程度就越高］；另外，记述统计学方法用来对数据集进行处理。比较多个数据的相似性是利用相关性来度量，基因表达的这种模式可以通过非监督算法（unsupervised methods）来描述，这种方法所得到的基因或样本的表达模式是通过分析基因表达数据的特征得到的。对于监督算法（supervised methods）而言，所要分析的样本（或者基因）和已经存在的已知类（例如正常组织 vs. 受病组织）有一定程度的联系，这样就可以利用基因表达数值，根据分析的类和已知类的相关性程度来预测受病样本。部分引用自 Brazma and Vilo（2000）

含有的字母代表样本名称。芯片技术主要是在同一个阵列里两个样本的杂交，所获得的基因表达值是比值或者相对亮度。

多数的基因表达数值是 Cy3 染料（绿色）和 Cy5 染料（红色）的亮度值的比值，其中每个信号的亮度被直接假设为每个基因的 mRNA 含量的比例。其他常用方法还包括利用单个样本和芯片杂交来得到数据值。这些方法多数用在含有放射性物质的 cDNA 芯片上，或者用在把寡聚核苷酸固定在模板上的方法制造的芯片中。这些芯片中所获取的基因表达的绝对数值将至少从两种实验条件下得到。每个基因根据这些绝对数值就可以得到相应的比值。

基因芯片表达数据（MGED）体系现在已经形成，它为芯片数据的分析、注释、交换和芯片数据的发布制定了标准，它的网站地址为 http://www.mged.org，由 MGED 集团发布，该网站包含许多重要的信息，其中包括芯片实验的标准化问题（Brazma，2001；Ball 等，2002a，2000b），还包含芯片实验标准化的最详细的细节（MIAME）（Brazma 等，2001），以及为能够对芯片数据统一保存、发布和相互交换而设计的关于基因芯片表达的通用语言（MAGE-ML）（Spellman 等，2002；Brazma 等，2000）。

我们将从三个方面（图 7-1）来描述芯片数据分析的问题。首先，数据归一化和预处理，这是让两个或者更多样本中的数据集能够相互比较的首要条件。其次，利用推测统计学（也叫假设检验，hypothesis testing）的方法分析，通过这种方法，可以对某些基因受到重要调控的可能性有综合的了解。最后，利用探测统计学（exploratory statistics）（也叫记述统计学）方法（包括聚类分析和主成分分析等算法）来检测复杂数据集中具有生物学意义的样本或基因表达模式。在一些芯片研究中也用分类方法根据基因表达谱来诊断一些物理实体（例如，癌细胞 vs. 调控细胞）。在芯片数据分析方面有许多非常优秀的综述（例如，Quackenbush，2001；Sherlock，2001；Dopazo 等，2001；Brazma 和 Vilo，2000 等文献）。

7.2　芯片数据分析：预处理

基因表达的改变可以反映出生物内在转录过程中某种相关性的选择，或者它们的表达改变是因为外在的多种多样的实验条件造成的。这些外在的人为的条件主要包括以下几个方面。

① 核苷受自身带有的荧光或射线的影响从而改变表达水平。

② 在芯片制造过程中，cDNA 固定在芯片上需要固体支持物，cDNA 受外在固体物质的干扰也会改变表达。

③ 在用荧光扫描（用来定量检测含有荧光染料的芯片表达值）或者用磷光计（用在基于放射性原理制造的芯片上）时造成基因的变异。

④ 技术原因，在生物样本的研究中，RNA（或 mRNA）的纯化或数量控制不够精确。例如，在复杂组织（如大脑）中，不同类型细胞中的基因表达是不同的，对它们分离却非常困难。

⑤ 在芯片实验过程中，RNA 的纯化、分离和杂交过程出现误差。例如，测定细胞系中在两个时间点的基因表达数量，研究人员纯化出来不同数量的 RNA，这就导致了芯片实验的人为误差。

⑥ 在洗涤清除芯片上面非特异性结合物质时也会出现误差。

Terrence Speed 研究团队和它的合作者已经开发了一些芯片杂交实验获得的数据归一化的方法，详细情况见 http://www.stat.Berkeley.edu/users/terry/zarray/Html/inder.html。许多团队已经探讨过了一些基本的数据归一化方法，这些方法包括，背景去除和全局均值归一化（background subtraction and global mean normalization）（Beissbarth 等，2000；Hegde 等，2000；Schuchhardt 等，2000；Liao 等，2000；SmidKoopman 等，2000；Sawitzki，2002）。其他开发的还有如何识别不同表达量基因的算法（Claverie，1999；Eickhoff 等，1999；Hilsenbeck 等，1999；Manduchi 等，2000；Wittes 和 Friedman，1999）等。

⑦ 信号检测过程中的误差。

数据预处理的主要目的就是尽最大可能去除在转录过程中因生物学内在相关改变的误差造成系统内在的偏差（Schuchhardt 等，2000）。而数据归一化的基本假设是在实验中，基因的平均表达量是不会改变的。

7.2.1 全局归一化

在本书中讨论的芯片数据归一化的问题并不是指使数据具有标准（高斯）分布，而是指通过归一化的问题校正两个或者更多的数据集中的数据，从而使这些数据集中含有的基因表达数据能够相互可以比较。

我们举例说明为什么对芯片数据的归一化是必需的：Cy3 和 Cy5 染料与 cDNA 结合具有不同的功效，如果不进行数据的归一化，就不可能精确地知道在含有这些染料的芯片中样本的相对表达值，从而在数据可比较的水平上，正确表达的基因所得的比值将和 1 相差很远；同时，基因芯片数据的归一化也使多个芯片实验的基因表达数据能够相互比较。因此，芯片数据的归一化问题无论在单通道（one-channel）还是双通道（two-channel）的芯片数据实验中都是必需的。

在芯片数据归一化中，首先需要测定背景颜色信号并且把它从每个基因表达信号中去除（Beissbarth 等，2000）。而背景信号的确定可以利用芯片中不含 cDNA 的空点来估计。这些背景信号可能在芯片表面是常量，也可能在局部是可变的（下面就讨论如何修正局部的背景信号的问题）。

大多数实验分析人员都是利用全局归一化的方法来处理芯片的原始颜色亮度数据，这样，芯片实验获得的基因表达数据比值的平均值就为 1。在芯片数据的归一化中，有个最基本的假设就是在检测的生物样本中，基因表达水平的平均值不会改变，因此，归一化后，其比值的平均值为 1 也就不难理解了。数据的全局归一化的方法可以用来分析双通道数据集（例如，Cy3 和 Cy5 标记的样本）和单通道数据集（例如，Affymetrix 芯片数据）。双通道数据集在处理过程中，可以分解为两个单通道的数据集，这样每个格点的颜色信号就可以被从这个通道中获得的特异性判断因子分开。对于需要归一化的两个或者更多的数据集来说，在单通道中（Cy5）所有基因表达测得的颜色亮度值都会乘以一个常量因子，这样，所有得到的红色和绿色亮度值就相等。例如，在绿通道中样本的表达均值是 10000 个单位（单位大小可以自己制定，只要是对所有的处理其标准都统一即可），而在红通道中样本的基因表达均值是 5000，那么在红色通道中每个基因表达值就会被乘以 2。如果所得到的数据不经过对数（lg）转换，比值的均值就为 1（在我们的举例中，10000/10000＝1）。如果数据经过对数转换（见下文），所得的均值就为 0。

其他全局归一化的方法也是可行的。一些实验分析人员把所有的芯片数据表达值归一化到在芯片上表达的管家基因（housekeeping genes）的数集中，这些基因包括 β-肌动蛋白和

3-磷酸脱氢酶甘油醛以及数十个其他基因。这样，在单个芯片中每个基因的表达值就可以被这些管家基因的平均表达值分开。利用这个方法的主要假设为：这些管家基因在芯片实验中，其表达值不受实验条件的影响。不过，在一些例子中，这样的假设是不正确的。而寻找好的管家基因的一个办法就是分析所有的候选管家基因在各种组织和条件下的表达量，如果某些基因的表达量无论在什么组织和条件下都不改变，或者改变幅度很小，那么这些基因就是比较好的管家基因。在一个研究项目中，研究人员测出了人类 19 个组织中大约 7000 个基因的表达，并且把这些数据对外公布，该数据库为人类基因表达（HuGE）检索数据库（Haverty 等，2002），在这个数据库中，包含 451 个管家基因，这些基因在人类的 19 个组织中的表达都是相同的，几乎没有什么改变。

> 人类基因表达（HuGE）检索数据库的网址为 http://www.HugeIndex.org。

7.2.2 散点分析

芯片数据分析方法中，其中最重要的可视性算法之一就是散点分析。在本节讨论中，所举的例子是芯片实验中两个样本基因表达数据值的比较。从图 7-2 中可以看到，多数数据点都落在散点分析图的 45°线上，但对于那些基因表达值相对于正常表达量高或低的数据就会偏离该线。这种散点分析图的算法能很快显示哪些基因在这个实验中受到了比较重要且非一般的调控。图 7-2 为散点分析图软件包所绘制的图表。

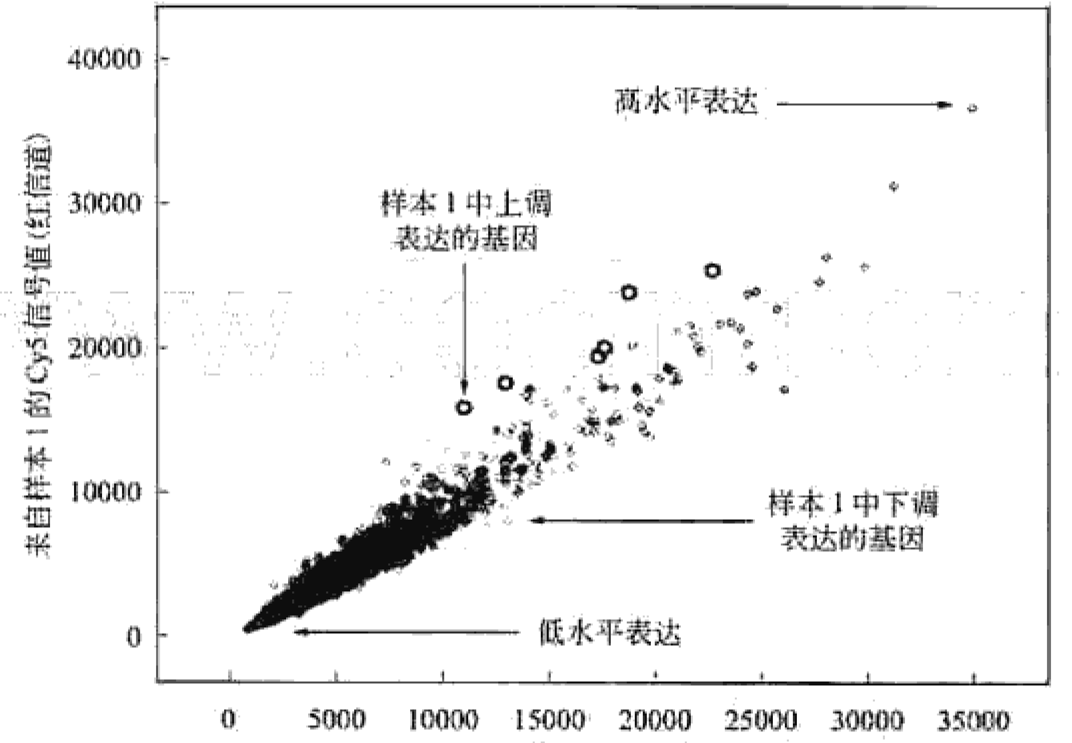


图 7-2　散点分析图为芯片实验所得的基因表达数据提供了一个最基本的分析方法。如上图，Cy5 信号（红染料）在 y 轴上，Cy3 信号（绿染料）在 x 轴上。在图里每个点代表一个基因，并且代表表达值低的基因或者空白信号的点都位于左下方，而代表表达量高的基因的点都位于右上方。在图中，45°线把数据点分成为两部分，在 45°线下方的点代表表达值低于它们正常表达值的基因，45°线上方的点代表表达值高于它们正常表达值的基因，并且越偏离 45°线，这种差别就越明显。直接利用上图的方法得到线性关系比把数据转换为对数所得到的线性关系好，不过把数据转换为对数也是很容易做到的。此图是利用微软的 Excel 制表得到的

在本节书中，将利用从 Chu 等（1998）的数据集来详细阐明散点分析图的问题。他们研究了在啤酒酵母（*Saccharomyces cerevisiae*）发芽的孢子发育过程中的基因表达的问题。该数据集可以很容易地以电子表格（框 7-1 和图 7-3）的形式下载下来。利用微软的 Excel 表格或者其他的各种图像处理软件包，该芯片表达数据就能很容易地转化为散点分析图表［图 7-3(b)］。而对于散点分析图（或者对于大多数芯片数据图表分析方法）来说，其最大的特点是在芯片实验中两种条件下的基因表达数据值的真实相关性可以通过对比表现出来。另外的一个特点为在散点分析图中颜色亮度较暗的点占主要部分。这就意味着在芯片实验中，大部分基因表达都处于较低的水平上，只有一小部分基因（见图 7-2 右上部分）是在较高的水平上表达。

框 7-1　芯片实验数据分析

芯片数据库可以让我们得到芯片实验的原始数据，这些数据包括芯片实验结果和基因表达的原始数据（表 6-6）。可以得到这些数据的一个地方就是斯坦福大学的芯片数据库。

- 点击进入网站 http://www.dnachip.org。
- 点击进入“已发布数据”的链接，找到 Chu 等人公布的一个数据集的研究。该数据集也可以通过进入网站 http://cmgm.stanford.edu/pbrown/sporulation/得到。
- 进入上面所述网站后，选择“additional figures and complete data set”，点击进入。
- 进入后，然后点击进入“SPO spreadsheet”。
- 当该页完全打开后，选择保存就可以完全把该数据集下载下来，把数据存入电脑硬盘里就可以了。
- 打开微软的 Excel 软件，然后从中打开“spospread.txt”文档。
- 在该电子表格中，第一列表示酵母开发可读框（ORFs）的标识符，以后的两列分别表示在 t_0 时刻基因在绿染料（Cy3）和红染料（Cy5）模式下的表达量，以后其他列表示在其他时刻基因表达数据以及其他一些数据分布（例如背景值等）。
- 选择第 2 列和第 4 列，利用微软的 Excel 软件作图功能中的散点图得到它们的散点分析图。该图见图 7-3，并且从图中可以看出数据特征。也可以把该数据输入到其他很多的软件中分析，这些软件包括 S-PLUS、Partek 和 SAM（一个微软 Excel 软件的插件，在本章中有介绍）。

散点分析方法的主要用途就是检测出在芯片实验的两种实验条件下受到不同调控的基因。不过，有些数据集利用散点分析的方法几乎找不到远离 45°线的点，几乎所有的基因表达数据都落在 45°线上及其附近［图 7-3(b)］。而我们最关注的却是那些远离 45°线的数据点，因为这些点代表着受到了重要调控的基因。

数据分析过程中，数据形式的转换有时对寻找重要的调控基因也是很有帮助的（图 7-2～图 7-5）。例如，可以把数轴上的数据从线性分布转换成对数分布［图 7-3(c)］。从图中可以看出，数据集在上面的分布大概呈在 45°线左右的对称分布。对于进行这样的转换，主要有两方面的原因。首先，通过转换，把数据点主要在图左下方分布［图 7-3(b)］改变到主要在图的中心位置分布，这样，数据集的特征就比较容易分析；其次，通过对数转换，基因受调控比正常表达改变的倍数更容易描述（例如，2 倍上调或者 1.5 倍下调等）。

数值转换中，还会对芯片的基因表达比值进行转换，例子如下：从一个芯片实验（表 7-1）中得到比值，基因表达的数值是从 $t=0$、1、2、3 四个时刻测得的，在 $t=1$ 的时刻基因表达值没有改变，而在 $t=2$ 的时刻基因上调表达 2 倍，在 $t=3$ 的时刻基因下调表达 2 倍。它们的原始数据为 1.0、2.0 和 0.5。2 倍上调和 2 倍下调拥有同样数量级的改变，只不过是相对于不同的方向。在原始的数据空间中，1 和 2 之间相差 +1.0，而 1 和 0.5（例如，上面

(a)

	A	B	C
1	ORF	t0 green	t0 red
2	YHR007C	8570	3643
3	YOL109W	7534	12218
4	YAL056W	1441	1043
5	YAL058W	2145	1740
6	YAL059W	1894	1576
7	YAL060W	7927	8770
8	YAL061W	5208	5664
9	YAL062W	8258	9527
10	YAR002W	2374	1838
11	YAR003W	2131	1397
12	YAR007C	2185	1553
13	YAR008W	1702	964
14	YAR009C	4848	4079
15	YAR010C	10550	9306
16	YAL001C	1530	1019
17	YAL002W	2302	1881

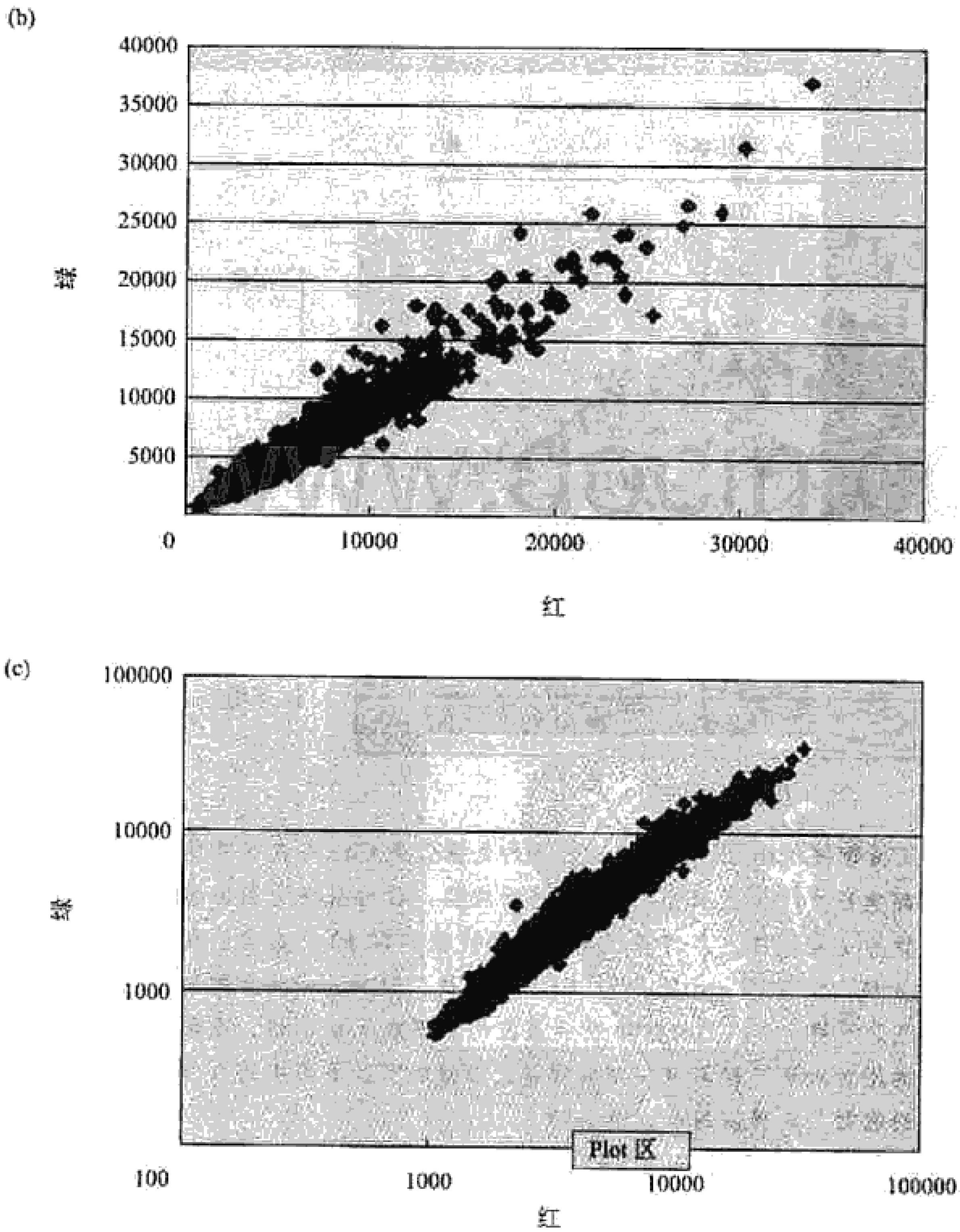


图 7-3 利用 Excel 软件得到的芯片数据的散点分布图。该数据集来自 Chu 等人（1998），可以从斯坦福大学芯片数据库中下载得到（网址 http://www.dnachip.org）。(a) 这个表格中，首列是酵母开放可读框的标识符，数据为啤酒酵母孢子形成时某时间段上基因表达变化值。该数据集包括 6118 个基因，绿色和红色分别代表酵母孢子形成过程中的不同实验样本。(b) 在得到该数据集后，该数据集可以在 Excel 软件中打开并且散点作图。本图所显示的为基因在 $t=0$ 时刻所得的数据作成的散点图。(c) 把数轴上的数据从线性形式转换为对数形式，同样对 $t=0$ 时刻的数据作图，我们可以看到数据点分布比 (b) 有明显的改变，得到的数据点的聚集程度甚至比 (b) 中左下方数据点的聚集程度都严重

表 7-1 从芯片实验中得到的比值数据

时间(t)	基因表达情况	原始的比值数据	log_2对数转换
0	基因的标准表达	1.0	0.0
1	相对标准表达无改变	1.0	0.0
2	两倍上调表达	2.0	1.0
3	两倍下调表达	0.5	−1.0

注：对基因表达数据的比值进行对数转换，所得数据比原始数据能更清楚地说明问题。

例子中 $t=1$、3 的时刻）之间却相差−0.5。而在对数空间中（例如，以 2 为底的对数），上面的问题就可以解决，经对数转换后，它们会呈现以 0 为基准的对称分布。

7.2.3 数据全局归一化中的局部归一化

在芯片实验中，许多人为因素对一些基因的表达改变是随机的，从而造成实验误差。对于这样的数据，可以通过局部数据归一化处理来减小实验误差，该方法对数据的处理具有偏向性，对信号颜色值并不是采取统一形式的处理。这里，我们继续利用 Chu 等的数据集。噪声的去除和数据的全局归一化如上节所述。同时，将数据进行对数转换，这样有利于寻找实验样本中的调控基因（图 7-4）。

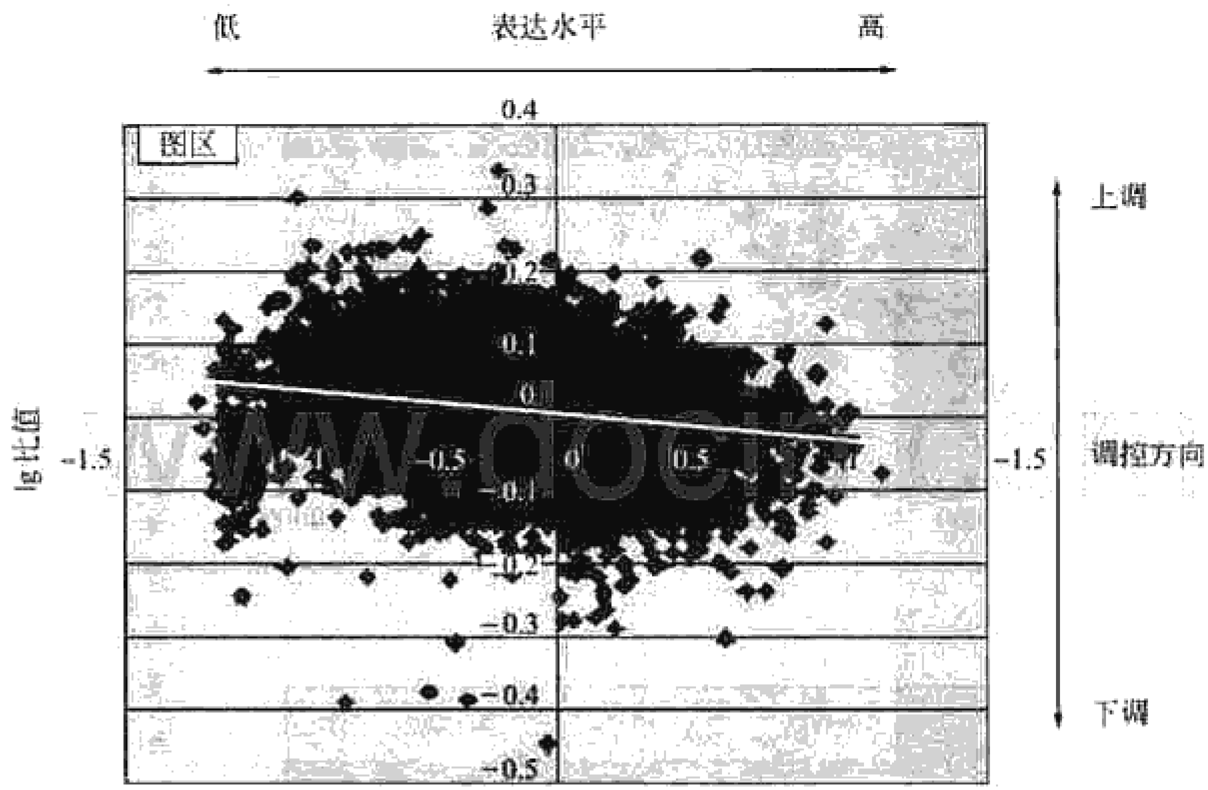


图 7-4 芯片实验所得数据转换后的散点图。利用的数据仍是 Chu 等人公布的数据，对基因表达颜色亮度值进行以 10 为底的对数转换，然后求得它们的算术平均值，该值就作为 x 轴的原点，而基因表达数据比值的以 10 为底的对数值就作为 y 轴上的数值，该轴相当于图 7-3（b）中 45°线。这样，x 轴上的数值表示的就是基因表达水平，y 轴上的数值表示基因的上调或下调调控。在图中，数据整体构成的图相对于 x 轴具有一定的倾斜，该倾斜可以用图中的白线反映出来。这种倾斜有时可以得以纠正（图 7-5）

在数据处理过程中，对基因表达颜色亮度值进行以 10 为底的对数转换，然后求得它们的算术平均值，该值就作为 x 轴的原点。这样，在这个 x 轴原点左边的所有基因都是低表达，而在 x 轴原点右边的所有值都高表达。在 y 轴中，散点所对应的值为芯片每个格点中

所得两种颜色亮度值比值的以10为底的对数值。利用这种格式很容易找到上调表达的基因（在 y 轴上数值较大的散点）和下调表达的基因（在 y 轴上数值较小的点）。对于 x 轴和 y 轴上数据转换的标准定义见框7-2。

在图7-4中，散点在 y 轴上的数值主要强调表达量的相对改变，散点在 x 轴上的数值表示基因表达的绝对值。这些数据构成的点图相对于 x 轴的倾斜和标准的散点图（图7-2和图7-3）相比仍然是比较清晰可辨的。这种倾斜有时候反应的是在实验过程中人为造成的误差改变，例如RNA受到染色体DNA或rRNA（这些核酸污染物可以结合到芯片的格点中）的污染等。另外导致人为误差的原因是芯片实验过程中利用了不等量的射线和荧光探针。

在图7-4中的数据点图的倾斜可以通过局部数据归一化改正（图7-5）（Colantuoni等，2002a，2002b），这种数据处理方式可以使基因表达比值在 x 轴得以平衡，这样，在需要相互比较的样本中所有基因表达的比值的均值就为0（在对数形式下）或者1（在原始数据比值形式下）。这种修正的方法主要是在数据点中插入一个局部弯曲的线，使这个线上下的点数对称。而新的 y 轴就包含一个芯片数据比值对数的校正值。该校正值是通过以下方法得到：首先对芯片实验中每个基因数据比值进行以10为底的对数转换，然后把它们相加，最后再减去局部颜色亮度值的均值，这样就得到了校正值（残差）。

在R统计语言中，"loess"函数用来计算局部基因表达比值的均值。芯片数据的全局和局部归一化可以在芯片数据标准化和归一化（SNOMAD）的网上来完成（Colantuoni等人，2002a，2002b），它可以从网上免费得到，网址为http://pevsnerlab.kennedykrieger.org。其他一些软件包[如基因转换软件（Gene Traffic software），网址为http://www.iobion.com；TIGR芯片分析系统（MIDAS），网址为http://www.tigr.org/software/tm4]都用了"loess"函数。

框7-2

利用散点分析图（如图7-2和图7-3），我们从图中芯片实验的两个样本（对照和实验条件）得到信息，它们的相似性远远多于它们的不同性。而我们对芯片实验数据分析，主要就是想得到它们的不同性。为了这个目的，必须对原始数据格式进行对数转换，同时调整 x 轴和 y 轴：对基因表达颜色亮度值进行以10（或2）为底的对数转换，然后求得它们的算术平均值，该值就作为 x 轴的原点，而基因表达数据比值的以10为底的对数值就作为 y 轴上的数值。如图7-4，芯片每个格点数值计算如下：

X＝lg(颜色亮度值的几何均值)

　＝lg(对照样本颜色亮度值×实验样本颜色亮度值)$^{0.5}$

　＝1/2[lg(对照样本颜色亮度值)＋lg(实验样本颜色亮度值)]

　＝以10为底的颜色亮度值的对数均值

Y＝lg(颜色亮度值比值)

　＝lg(实验样本颜色亮度值/对照样本颜色亮度值)

　＝lg(实验样本颜色亮度值)－lg(对照样本颜色亮度值)

　＝lg(基因表达比值)

从计算可以进一步明白，由对照样本和实验样本中的基因表达颜色亮度值进行以10（或2）为底的对数转换，然后求得它们的算术平均值，该值就作为 x 轴的原点，所以 x 轴上的数据就表示基因表达的水平；而对照样本和实验样本中基因表达数据比值的以10（或2）为底的对数值就作为 y 轴上的数值。

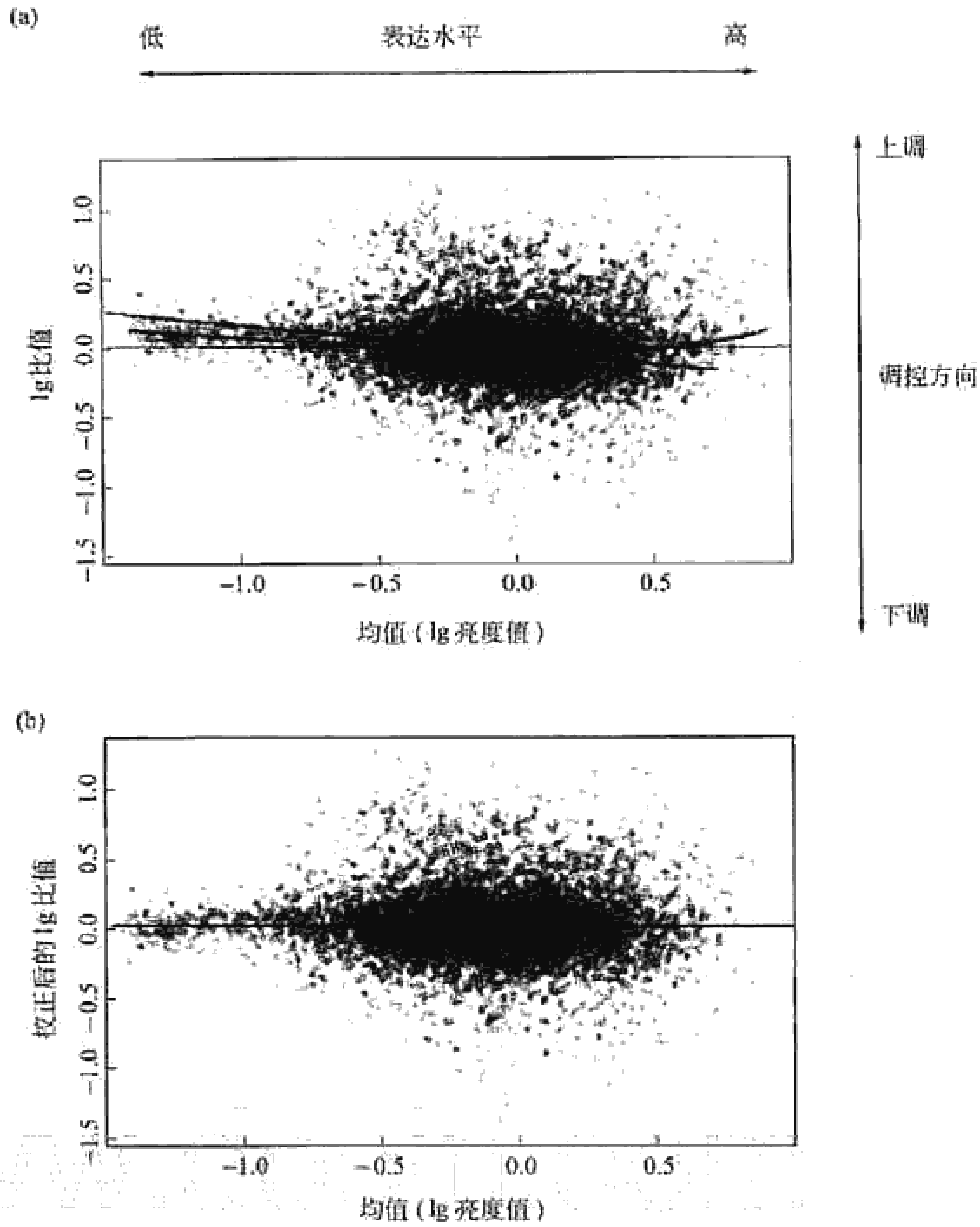


图 7-5 芯片实验数据散点图的局部噪声修正。(a) 如同图 7-4，对基因表达颜色亮度值进行以10为底的对数转换，然后求得它们的算术平均值，该值就作为 x 轴的原点，而基因表达数据比值的以10为底的对数值就作为 y 轴上的数值。从图中可以看到，数据点所构成的图相对于 x 轴的倾斜用图中的一条直线表示出来，而这种倾斜可以通过利用计算得到的局部调整倾斜度的线来校正（图中红色的线）。(b) 在计算最优的调整倾斜度的线时，每个数据点相对于该线的距离（残差距离）用来调整该线的位置。所有的这些转换可以见SNOMAD，网址为 http://pevsnerlab.kennedykrieger.org/snomad.htm

7.3 芯片数据分析：推测统计学方法

在芯片实验后，怎样才能辨别出哪些基因所受到的调控是重要的呢？其中，一种方法是通过计算对照和实验样本下基因表达数据的比值，然后对它们的比值从大到小进行排序。进行完排序后，人为地定义一个阈值（例如，某些基因的调控表达值2倍地高于或低于正常表达时的表达值），定义了这个阈值后，表达高于该值的基因就是我们感兴趣的基因，其他的基因就舍弃掉。不过，这里面的最大问题就是阈值是我们人为的定义的，而我们并不知道该值定义为多少才是最合适的，也就是说，根据该阈值舍弃掉的基因也许有一些反而是比较重要的基因，或者根据此阈值选择的一些基因反而是不重要的基因。这种标准应用在一些芯片实验中，计算得到的最终结果往往是没有基因或者很少的基因被选择出来，而与此相对的是，在另外一些芯片实验中，又有成千上万的基因根据这种标准被选择出来。另外，在芯片实验中，如果背景信号噪声是100个单位（人为定义的），而基因在实验和对照样本中表达水平分别为300个单位和200个单位，在去除背景噪声后，表达水平变为200个单位和100个单位，这样可以计

算出基因表达比值为2，表明这个基因受到了2倍于正常表达水平的调控表达。根据上面定义的判断标准，这个基因就算是比较重要的基因。不过，这个基因表达的绝对值和背景信号噪声非常接近，去除噪声信号后，得到的200个表达单位和100个表达单位就可能与真实值相差很大，误差也会很大，据此判断该基因在实验样本中受到2倍于正常表达水平的调控表达是不准确的。而比较能让人信服地判断一个基因受到了2倍上调表达的例子可为，在背景信号噪声为100个单位的情况下，在去除信号噪声后，实验样本和对照样本的表达水平分别为10000个单位和5000个单位，这样就可以说该基因受到了比较重要的调控。

把芯片数据的表达值转换为比值是非常重要的，因为从比值中可以很快找到哪些基因是比较重要的。不过，有了这些比值也代表一劳永逸了，因为并不能把这些比值转换成特定的概率值，而通过这些概率值则可以检验我们假设的某些特定基因受到重要调控的可能程度。因此，许多研究小组只是把表达比值作为许多判断标准中的一种来用于芯片数据的分析。例如，Iyer等（1999）年研究了人类纤维原细胞对免疫血清的转录反应，并且从中选择了表达比值＞2.2的基因用于后续的聚类研究（见下面的讨论）。

另外选择受到重要调控基因的一种方法是对芯片实验数据比值从大到小排序，选择前5%的基因作为后续研究的对象。例如，在某个芯片实验中，共有10000个基因被研究，那么我们就可以根据把表达比值排序在前500的那些基因选择出来，认为它们就是受到了重要调控的基因。这个方法也存在着难以解决的问题，该方法没有对在不同的对照和实验样本中基因的表达均值差异量度做出限制。然而，根据基因表达比值进行排序，并且根据排序找到一些在表达水平上与正常表达值差异较大的基因总是可能的。因而，这种方法在实际操作中是没有用的。

推测统计学方法对芯片数据的分析目的就是做假设检验，即假设在两种或两种以上实验条件下，某些特定的基因受到了重要的调控，而推测统计学方法的应用，就是用来检验这种假设的可靠程度。在假设中，首先定义H_0，即在两种或两种以上样本的芯片实验中，所研究的基因在表达量上与正常表达时的表达量没有显著的差异性；与H_0相对的就是H_1，即在两种或两种以上样本的芯片实验中，基因的表达有显著的差异性。我们利用推测统计学方法来完成假设检验，该方法是利用目标函数，根据基因表达值首先计算，然后通过阈值来选择基因。我们将根据推测统计学的结果来选择H_0或者H_1。在科学计算中，选择拒绝H_0的概率也有一个重要的标准，为显著性水平α，一般其概率值都设为$P<0.05$。

> 假如在两个样本中共有100个基因的表达量，并且样本之间的基因表达量没有显著的差异性，你就想通过（100）（0.05）＝5的方法得到5个重要的调控基因。不过，假设这些检验都是统计上独立的，那么得到至少一个比较重要结果的概率为：$1-0.95^{100}=0.994$（见Olshen和Jain，2002）。为此原因，利用此方法需要校正。

芯片实验数据能否应用统计检验的方法关键在于芯片实验的设计环节。以一个标准的芯片表达实验为例，该实验测得10个细胞系的基因表达值，其中5个细胞系来源于病人，属于受病细胞系（实验样本），另外5个细胞系来源于健康人群（对照样本）。可以根据芯片实验中基因的表达值计算出它们的均值和标准差。利用t检验方法来检验H_0，看这两种样本中基因的表达值是否存在显著的差异性。t检验方法为，首先计算基因在对照样本表达值的均值（x_1）和

实验样本表达值的均值（x_2），然后估计它们的偏差（σ），最后用x_1和x_2的差值与σ相除，公式为：

$$t=\frac{x_1-x_2}{\sigma} \tag{7-1}$$

t-检验方法最基本的用途就是测芯片实验中的信噪比，通过信号（样本均值的差值）和噪声（两个实验样本中的偏差）的比值得到。从t统计表中可以计算出P值，根据P值就可以决定接受还是舍弃H_0，如果接受H_0，就说明在这两种人群中，所检验的基因表达水平没有显著差异性，如果舍弃H_0，也就是接受H_1，即基因表达水平差异性显著。而t检验方法的假设就是基因表达值是呈正态分布的。

> 许多商业性芯片分析软件包都含有芯片数据的假设检验，并且大多数包含对显著性水平α值的校正和选择功能，这些软件同时也包含对数据变量的可视化分析工具，如Partek Pro（网址为http://www.partek.com）、GeneSpring（网址为http://www.sigenetics.com）、GeneSight和Genedirector（网址为http://www.biodiscovery.com）、Gene Traffic（网址为http://www.iobion.com），以及Spotfire（网址为http://spotfire.com）。

在芯片数据分析中，对显著性水平α的校正是必需的，因为数据的分析有多重检验的问题。例如，在芯片实验中得到了10000个基因的表达值，如果假定大约其中5%的基因受到了比较重要的调控，那么显著性水平α就可以为$P<0.05$的水平。不过，如果假定在这个实验中，有一个特殊的基因受到了非常重要的调控，那么显著性水平α值就需要做出适当的调整。但是，对于10000个芯片数据来说，为了适应做出的数以千计的假设，显著性水平α值就要有相应的校正，这样的工作是比较烦琐的。为了改变这种情况，许多研究人员利用Bonferroni校正法，此方法是用显著性水平α值除以芯片实验所得的总数据值（例如，$P<0.05/10000$被设为判断显著性的新标准）。

在统计检验中，有许多种方法都能用来处理芯片实验数据（如Olshen和Jain，2002），其中一些统计方法在表7-2中已经列出。所有这些检验方法的目标都是要获得P值，然后根据P值来判断基因受到特别调控的可能性。对于多于两种实验条件下的样本（例如，芯片实验中的多时间点分析或者多种药物对基因表达的影响等），利用向量分析（ANOVA）方法比t检验方法更加适合。利用向量分析的方法，可以计算出多因素有相同分布可能性的概率。

表7-2 芯片实验的统计检验

样本类型	参数检验	非参数检验
单样本	单样本t-检验(one-sample t-test)	威尔科克森检验(Wilcoxon test)
不配对双样本比较	不配对t-检验（unpaired t-test）	Man-Whitney检验(Man-whitney test)
配对双样本比较	配对t-检验（parid $t$test）	威尔科克森检验(Wilcoxon test)
3个或更多不匹配样本比较	单向变量分析法(one-way ANOVA)	Kruskal-Wallis检验
3个或3个以上匹配样本比较	反复测定分析法(repeated-measures ANOVA)	Friedman检验

假设检验分为参数检验和非参数检验。参数检验用在样本的数据集为正态（高斯）分布的情况下。最基本的参数检验包括t-检验和向量分析法。非参数检验并不对

样本数据的分布用任何假设前提，它们只对得到结果向量（基因表达值）进行从高到低的分类排列，并对这些分类进行分析。非参数检验，例如 Man-Whitney 检验和威尔科克森检验（Wilcoxon test），它们在芯片数据分析上被使用的次数远少于参数检验。

在芯片数据分析中，利用何种假设检验的方法也要根据芯片实验的样本特征来决定。一些芯片实验的样本可见图 7-6。对于两对象问题的芯片实验设计［图 7-6(a)］，

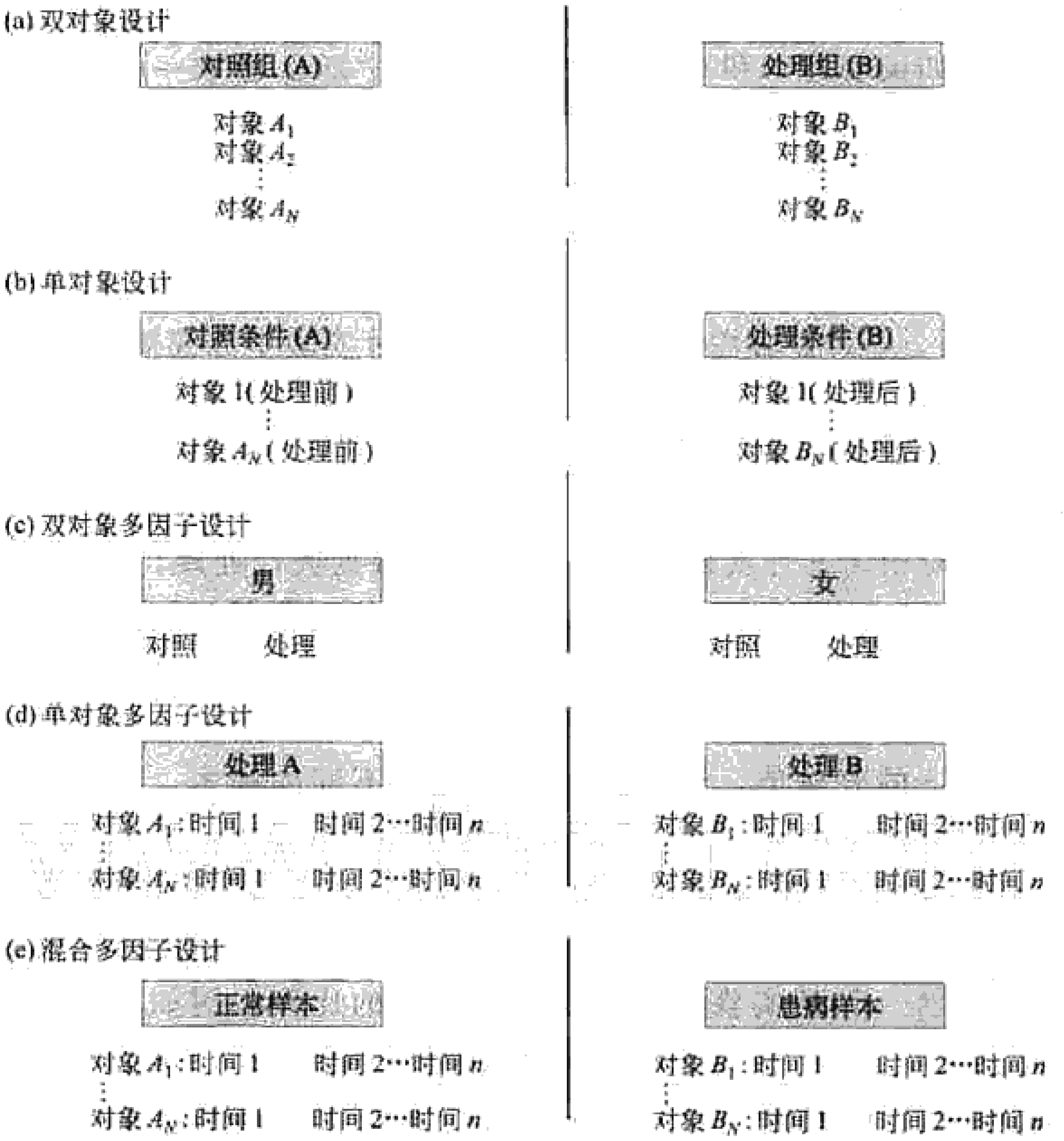


图 7-6 芯片实验设计和基因表达谱模式举例。在多数芯片实验中，设计人员都是为了想最终得到比较重要的有生物学意义的基因，这些基因在不同组织（正常组织对染病组织，或大脑中的基因对肝脏中的基因）、不同时间点或不同的药物处理中的表达量差别显著，可以通过统计检验的方法得到。(a) 在两种对象下的芯片实验，必须对一些因素（如年龄、性别和体重等）进行控制。(b) 单对象条件下，没有遗传变量的因素对实验的影响可以用来测定基因在经过某些处理前后的表达情况。(c) 两种对象、两种处理方式的芯片实验，不但可以测得基因在对照和实验样本下表达量的不同，而且可以测得基因在其他特征（如性别）下表达量的不同。(d) 单对象多因子设计是用来研究两种处理在不同时间上的基因表达。(e) 混合因子设计，双对象（如正常组织对染病组织）和单对象多因子（如多时间点上基因表达测定）

芯片包含的样本为两个。Golub等（1999）测得了急性白血病的两个亚型样本的基因表达值。在这个实验设计中，控制两个实验样本中病人的特征是必须的，例如不同年龄、性别或体重。对于单对象芯片实验设计［图7-6(b)］，配对检验将用来检验在配对样本中测得的两个样本数据均值的差异性。关于这方面的研究，Perou等（2000）测得了乳腺癌在用药前后组织切片的基因表达。这种情况下，一些相关的协调变量（如年龄、性别）受内部的调控。

在芯片实验中，如何才能确定统计检验得到的概率值不是偶然因素，或者说，不是通过基因表达的随机改变得到的。利用交叉检验的方法可以解决这个问题。交叉检验是在每个样本（如染病样本和对照样本）的数据中随机选取一些数据组成新的测试样本，并且利用与分析实验和对照样本数据同样的统计检验的方法得到 P 值。在芯片数据分析中，可以多做一些交叉检验（如100～1000次），如果分析对照和实验样本得到的 P 值都比交叉检验所得的 P 值小，那么 H_0 就被拒绝，也就是说，样本中基因表达的差异性显著。

芯片显著性分析。芯片显著性分析（SAM）是芯片实验中寻找被重要调控基因的一个方法（Tusher等，2001），SAM是基于芯片实验中基因表达值的改变（标准差）对每个基因打分，根据分数的多少来判断基因的重要程度。

利用SAM可以得到许多有用的特征，并且操作比较简单，利用微软Excel插件就可以完成。它可以对很多芯片实验设计方法得到的数据进行处理，如图7-6中包含的数据。用户在操作SAM之前，首先要对芯片数据进行归一化并且把数据分组［这些处理都可以用微软的Excel和其他软件（如SNOMAD）来完成］。SAM可以处理芯片实验的原始数据类型或者经过各种转换后的数据类型。样本的数据矩阵中每一行是一个基因在各种实验条件下的表达值，而每一列对应的是不同的样本。SAM利用了改进的 t 检验（式7-1）来检验 H_0（见Tusher等，2001）。

SAM的一个主要特征就是它能提供错误发现率的信息，错误发现率是因偶然因素而被选择出来的基因占所有选择出来基因的比例。用户可以调整实验参数来控制假阳性的比例。例如，在某类型芯片实验中，利用 t-检验得到的每100个受到重要调控的基因中有10个是假阳性的，那么这个实验的错误发现率就是10%。用户可以在数据分析时降低假阳性率（所要付出的代价是许多真阳性的基因被丢掉），或者为了得到更多真阳性的基因，用户可以提高假阳性率（而付出的代价就是选择了更多的假阳性基因）。SAM算法可以计算出 q 值，即一个基因被指为受到重要调控基因的最低错误发现率。

> 可以从网站 http://www-stat.stanford.edu/~tibs/SAM/上得到SAM软件。并且，在做交叉检验时，需要做足够多的交叉实验（≥100个）和足够多的样本（例如，每个实验条件下最少得5个样本），这样才能得到比较真实的值。如果样本太少，得到的值就没有鲁棒性。

图7-7显示了SAM算法得到的输出结果。根据 t-检验得到的基因列在表中，并且作图［图7-7(b)］显示（把通过检验统计得到的基因分类，并作图显示观测到的基因数和预期的差别）。该图（q-q 图）有效地显示出了受到重要调控的基因。在SAM分析过程中，交叉检验仍然需要，利用交叉检验可以评定有表达量的基因的重要性，方法为在样本水平（如对照和实验条件下）下随机对每个基因进行100次或者更多次的 t-统计。

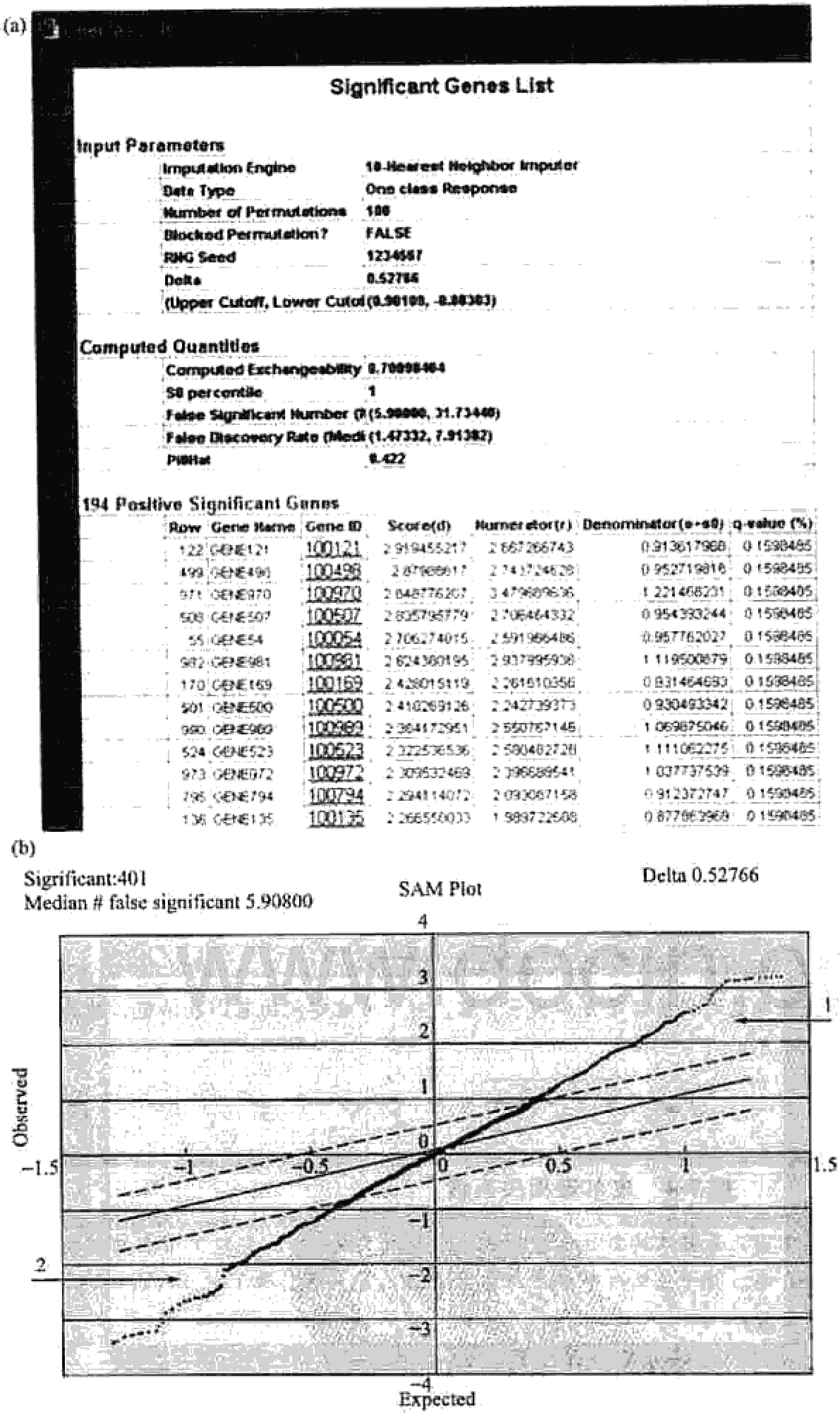


图 7-7　SAM 是微软 Excel 的一个插件，利用改进的 t 统计方法来寻找受到重要调控的基因。利用 SAM 分析，输入的数据是基因在不同条件下（如对照和实验）得到的表达数据矩阵，用户可以在假阳性率允许的范围内选择参数 Δ 来做为阈值。用户也可以选择以倍数改变值作为参数来决定阈值。(a) 输出结果为受到重要调控基因的列表。列表中分数 d 是对每个基因的 t-统计值；表格中分子表示两个相互比较的基因表达平均值的差，分母表示分子标准偏差的估计值；q 值为假阳性率。(b) 图中表示了预期值和观察值之间的关系。在图中，上调调控值比较大的基因（箭头 1）和下调调控值较大的基因（箭头 2）都用了不同的颜色区分开来，与表 (a) 中颜色所表示的含义相同

7.4　芯片数据分析：记述统计学方法

在芯片实验中，它的一个最基本的特征就是通过实验产生了大量的数据，基因表达值的数据远远大于样本量。而我们怎样才能分析得到在芯片实验中，利用 10 个细胞系中 10000 个基因表达所获得的结果的含义？我们可以把 10000 个基因的每个基因表达值转换为在

10000维空间的每个空间点值。而在我们的概念里，我们并不能完全理解并做出高维空间图。所以，我们必须利用数学方法把高维的数据空间降为低维数据空间。

关于数据分析的比较优秀的回顾性文章可见 Raychaudhuri 等（2001）和 Quackenbush（2001）。

对于有很多向量形成的向量空间，可以称之为"向量诅咒"（curse of dimensionality），解决这类问题需要利用数学方法。在高维空间中，任何两点之间的距离都很大并且它们的值都近似相等。利用记述统计学方法分析芯片的高维数据是比较有效的方法。记述统计学方法并不对问题提出假设，因此用这种方法也就不用测定显著性的结果。利用记述统计学方法分析数据集的目的是从中找到具有生物学意义的模式。例如用聚类树算法可以得到基因（或者样本）是怎样聚成某些有意义的类的。在做完聚类后，其中的某些基因又可以再进一步做假设检验。

对于记述统计学主要的流程可见图7-1。在记述统计学中，每个方法的目的都是为了把多维空间向量降到低维，从而能对单条件下的芯片实验或者多条件下芯片实验的基因表达或样本的分类进行分析。在每个例子中，我们首先都从数据矩阵分析开始（每行代表基因表达量，每列代表样本类型）。然后利用全局或局部归一化对数据预处理。然后定义某种参数（或者定义某种距离）作为描述样本点之间相似性的判别依据。

描述基因表达数据之间距离两个最常用的参数是欧氏距离（Euclidean distance）（框7-3）和皮尔逊相关系数（Person coefficient of correlation）（框7-4）。许多分析芯片数据的软件包在分析数据之前都必须设定距离参数，利用欧氏距离、皮尔逊相关系数或其他的距离函数来描述基因表达值之间的相关程度。

框 7-3 欧氏距离

欧氏距离（d_{12}）是描述在三维空间（坐标为 x_1、x_2、x_3和 y_1、y_2、y_3）中两个点之间的距离，公式如下：

$$d_{12}=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+(x_3-y_3)^2} \tag{7-2}$$

欧氏距离主要是表示两个特征向量之间差值平方和的平方根。对于 n 维空间的基因表示值，欧氏距离的定义可以如下：

$$d=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{7-3}$$

框 7-4 皮尔逊相关系数

在定义基因表达数据相关性方面，最基本的参数也许就是皮尔逊相关系数 r。它经常用在构建树的系统中，例如聚类（见下面讨论）等。对于任意两个多维向量 $X=\{X_1,X_2\cdots X_N\}$ 和 $Y=\{Y_1,Y_2\cdots Y_N\}$ 来说，它们的皮尔逊相关系数 r 定义为：

$$r=\frac{\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma_x}\times\frac{Y_i-\overline{Y}}{\sigma_y}\right)}{N-1} \tag{7-4}$$

式中，$\overline{X}$ 用来表示向量 X 的均值；σ_x 表示这些向量 X 所有值的标准差。

在散点图中，r 可以用来描述任何一条线适合这些值的程度。皮尔逊相关系数 r 的数值永远在 +1（两个多维向量完全相关）、0（完全不相关的数据集）和 −1（两个多维向量完全负相关）之间。

为了详细说明记述统计学方法，我们利用了 Chu 等（1998）的数据集的一小部分举例。在去除背景噪声和数据全局归一化后，我们得到了一个含有20个基因（行）和3个时间点（列）的数据矩阵［图7-8(a)］。在这个矩阵中，数据都为基因表达比值的对数形式。我们

(a)

	1	2	3
	log2.t0	log2.t0.5	log2.t2
1	−0.40	−0.91	−1.60
2	−0.99	−0.07	−0.83
3	−0.22	−0.49	−0.28
4	−0.31	−0.01	−0.09
5	−0.48	1.31	0.36
6	−0.38	0.35	0.60
7	−0.41	−0.49	−0.54
8	−0.46	−2.72	−3.16
9	−0.15	0.06	0.13
10	0.12	−0.67	−0.77
11	−0.03	−1.87	−2.58
12	−0.31	0.02	−1.64
13	−0.06	−0.22	0.17
14	−0.03	−0.23	0.02
15	−0.12	0.11	−0.01
16	−0.21	−0.66	−0.30
17	−0.40	1.66	1.13
18	−0.58	0.25	0.72
19	−0.77	−0.05	1.11
20	−0.28	0.43	−0.57

(b)

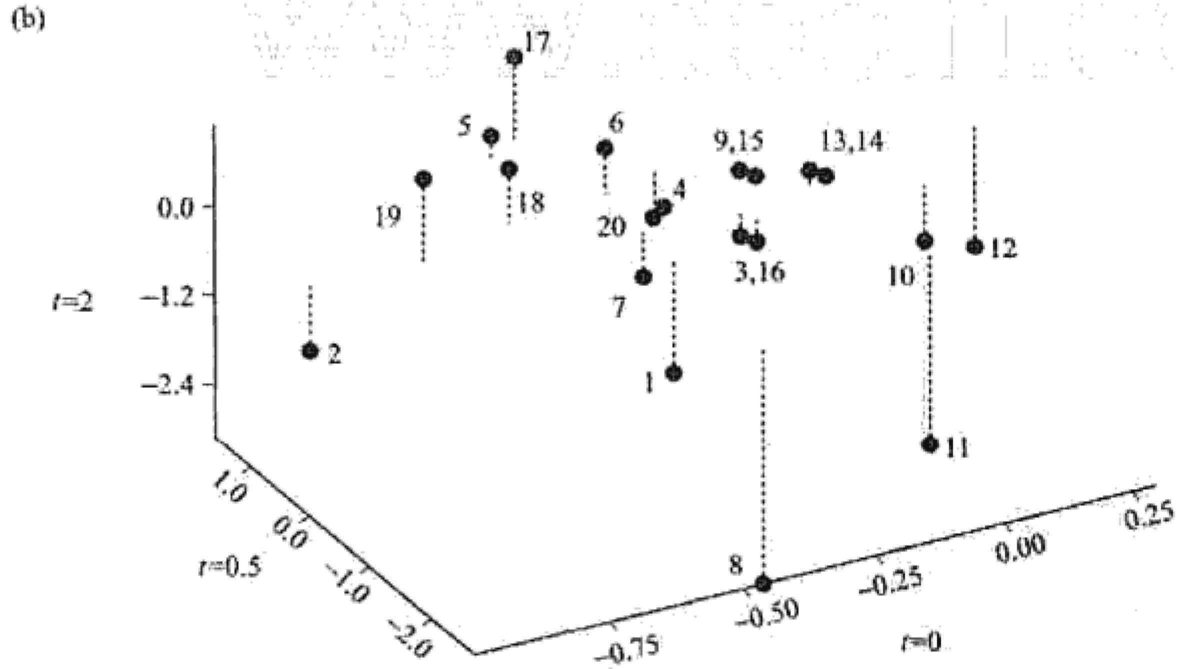


图 7-8　记述统计学方法分析芯片表达值的矩阵。数据集是某些基因在一个时间段上的表达值（Chu 等，1998），它可以转换为 Excel 格式（框 7-1），该矩阵中，背景噪声已经去除，对数据进行了全局归一化，并且把三个时间点（$t=0, 0.5, 2$）上的表达比值转换为以 2 为底的对数形式。在图（a）中，基因表达比值对数值在前 20 位的基因被列举出来。这些数据集将用来分析多种记述统计学方法，并且这些数据利用 S-PLUS 制图系统（insightful 公司）在三维图上显示出来。在三维图上，三个坐标代表了数据的三个时间点。而对于有很多数据（如上千个）的情况下，利用作图法是不实际的。但对于比较小的数据集，利用作图法就很容易发现受到特别调控的基因，例如基因 8 和基因 2

可以用三维图像把它们在图上显示出来［图 7-8(b)］。从图中，我们能看到一些点与其他的点偏离比较远，例如基因 8 和基因 11。然后，在从图 7-8(a) 中找到这两个基因，我们发现，它们在第三个时间点上［图 7-8(a)，最右列］基因表达量都相当大。而对于含有 6119 个基因表达值的数据集来说，我们利用如上三维作图的方法来寻找受到重要调控的基因是不实际的。同样，利用三维作图的方法来寻找数据集的不同模式（例如，寻找在三个时间点上基因表达向量的相对分布）也是困难的，因为点数过多时，就很难通过作图发现单个或其中

一些数据点的特征。

在要讲的记述统计学方法中，首先要说的是非监督算法（图 7-1）。非监督算法中，关于基因或者样本不必做检验假设，对于数据集的模式分类是根据基因表达活动的相关性来判断的。可以通过预先已知类的数据集对非监督算法进行检验，看它分类的准确程度。

7.4.1 芯片数据的层级聚类分析

对于单个芯片实验，并且只有两个样本数据可以用散点图来分析，并且受调控基因的显著性可以用 t-检验获得。多个芯片实验的数据可以用多个散点图分析，另外，也可以对它们含有的基因或样本进行聚类分析。在芯片实验中，聚类方法是寻找具有相同表达模式基因的一个很普通的工具。对基因的聚类可以通过聚类树、样本或者既根据聚类树又根据样本来完成。

聚类是通过两个对象之间的距离值来描述的，并且可以通过散点图或者系统树图［例如用来分析系统发育（第 11 章）或芯片数据的］来描绘出来。聚类的主要目标就是把两个对象数据值转换为相似度（或者距离），从而确定这两个对象之间的相关性程度。被聚类算法归结为同一类的点具有较大的相关性，而不在同一类的点相关程度就比较小。在聚类中，基于欧氏距离的距离矩阵是比较常用的。

聚类算法有很多种，其中在芯片数据分析中最常用的是层级聚类算法，层级聚类的结果是输出是一个系统发生树（在下文将要举一个非层级聚类方法：k 近邻聚类算法）。层级聚类算法可以用收敛（或者发散）方法来实现（图 7-9）。在每个例子中，输出的结果都是用树图来描述对象之间的相关性。在发散方法中，算法的第一步是所有的数据都属于 1 类（$k=1$）。在随后的计算过程中，根据参数再把每一类确定出来，直到达到 n 类为止。在收敛算法中，所有的被研究对象都是独立的。也就是说在每聚类之前有 n 类，每个对象自己形成一个孤立的类。然后对两两相似的类进行合并，直到剩下 1 类为止。

在聚类算法中，类之间的距离一般定义为一个类中所有的点和另一个类中所有点的距离的平均值，这种定义方法称为平均联接法，它一般用在求不加权重的两个类的平均值（UPGMA）。第 11 章将讨论 UMGMA 的具体实施步骤。在定义了以聚类树为基础的聚类算法后，其他一些聚类算法也在 Quackenbush 中提出（2001）。

收敛聚类算法和发散聚类算法分析相同的数据集会得到比较相似的结果。我们可以用前面提到的比较小的数据集（图 7-8），该数据集有 20 个基因和 3 个时间点。用该数据集利用收敛算法和发散算法得到两个聚类树（图 7-10）。对于每个聚类树来说，y 轴（高度）代表了不相似度。对基因 8 和基因 11，我们先前在图 7-8（b）中已分析过，它们和其他基因点偏离比较远，在图 7-10 中可以看到，它们的分支高度比较大。在聚类树中，x 轴代表的是基因（或者样本）的均匀空间，它们位置的显著性代表了它们属于某一类的显著性。从图 7-10 中可以看到，两个聚类树的整体拓扑结构很相似，但也有几个基因在收敛聚类树和发散聚类树中位置不相同（图 7-10 箭头所示）。总的来说，用不同的记述统计学方法分析同一个数据集，所得结果总有些许差别。那么，如何确定这些分析结果中哪种结果有更高的可信度？其中的一个方法就是需要在芯片实验中多做些独立的重复；另外，也可以通过生物学知识来检验

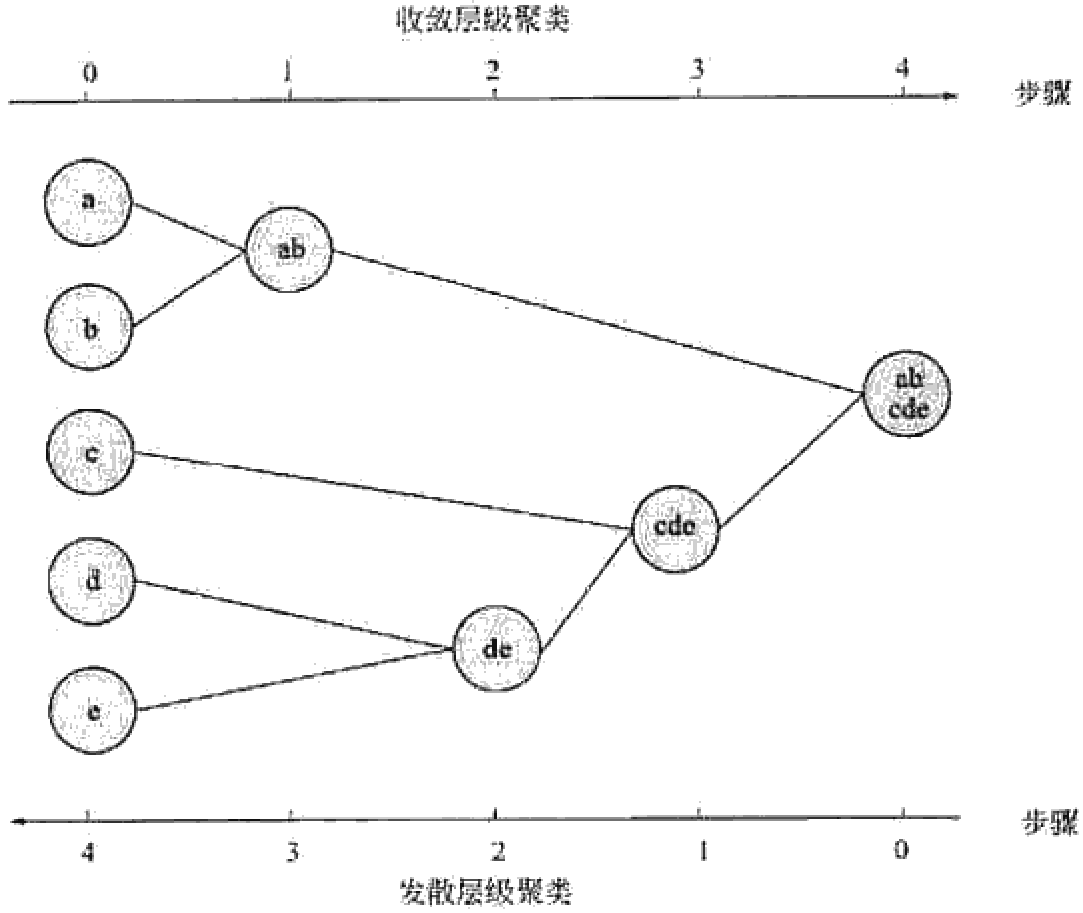


图 7-9 层级聚类算法的两种主要方式——收敛和发散。在收敛聚类算法中，每个数据点（基因或者样本，在图中用 a～e 表示）都是独立的（第 0 步）。在聚类时，其中的两个最相关的点被聚为一类（图中见圈 ab，此为第 1 步），而所用数据点之间的相关性用欧氏距离参数来确定。去除上两个点后，最相关的两个点聚为下一类（de，第 2 步）。如此继续下去（第 3 步，第 4 步）直到所有的点被聚为一类（abcde），这就是收敛聚类。而图中，形成这种聚类关系的图表就叫聚类树。而发散聚类算法与收敛聚类算法的步骤正好相反。在聚类开始，所有的数据点被归为一类（第 0 步，abcde）。然后，在该类中，与其他的最不相关的一个最该类中分离出来。然后分离次最不相关的。如此继续下去直到所有的对象都被分开。与收敛算法一样，发散算法可以得到一个聚类树。在实际应用过程，收敛聚类算法和发散聚类算法会得到相似的聚类树。此图来自 Kaufman 和 Rousseeuw（1990），授权使用

这些聚类树。如果基因 1 和基因 12 [在图 7-10(a) 中被聚类在一起，而在图 7-10(b) 中没有聚类在一起] 都是编码细胞因子（cytokines）的基因，那么就会认为收敛聚类算法得到的结果更可信。而对于这些结果的进一步检验可以用主成分分析算法（见下文）。

许多软件都可以用于芯片数据的聚类分析。在图 7-8 和图 7-9 中用了 S-PLUS（insightful 公司）。在芯片数据聚类软件中，其中最重要的一个软件就是 Cluster 和它的相关树可视化软件 TreeView。

Cluster 是由 Michael Eisen 开发，并且可以通过他的网站得到，网址为 http://rana.lbl.gov/。

这种软件（其他相似的软件）的输入数据格式为包含基因的表达值和样本量的表格 [图 7-11(a)]。如上面所述，数据可以是对数转换形式。数据的归一化可以使每行或者（和）每列向量数据值的平方和等于 1。数据过滤器可以把一些基因去除掉，主要是因为最大或最小值超过了阈值。在 Cluster 中，距离参数用的是皮尔逊相

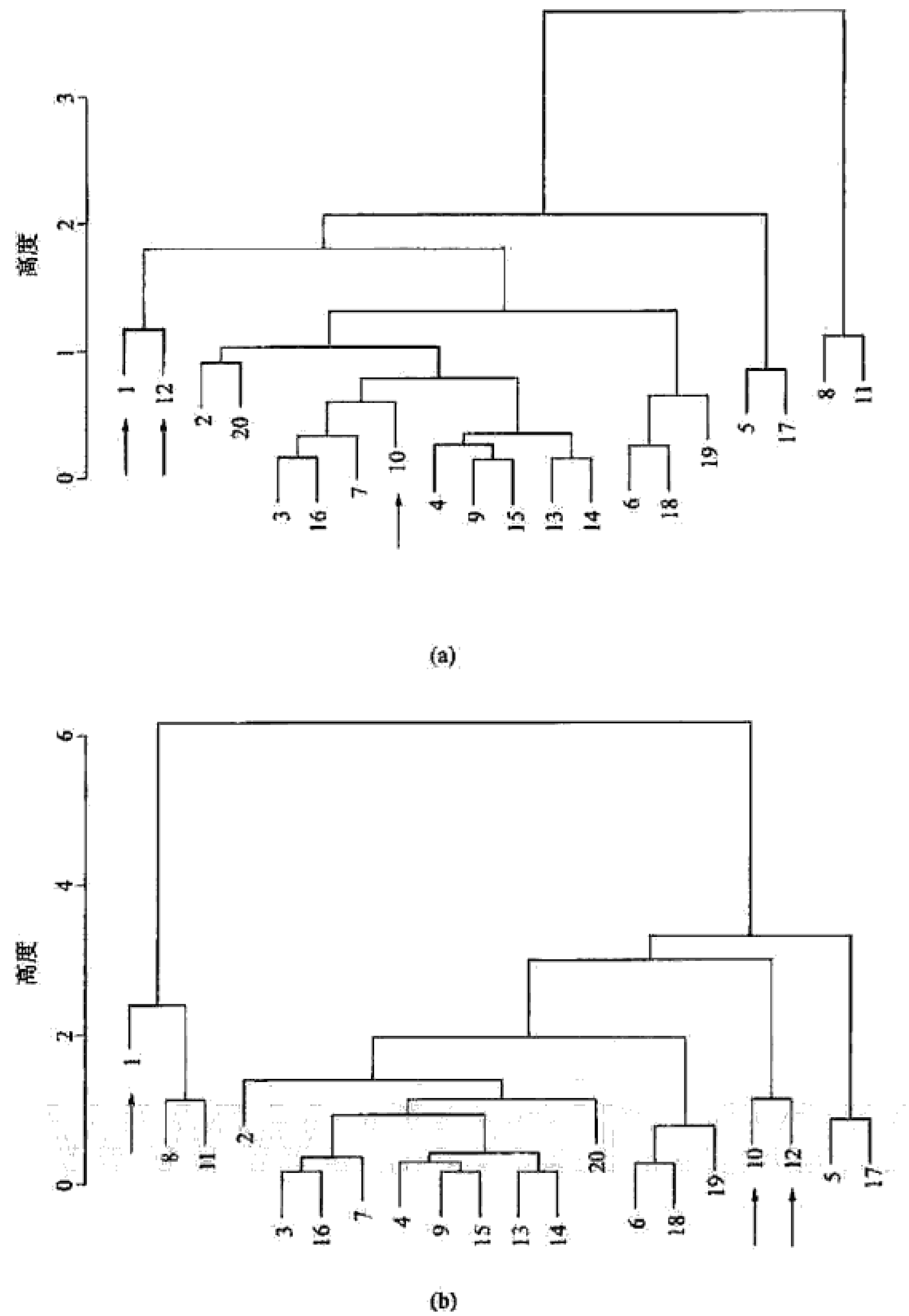


图 7-10 (a) 芯片数据的收敛层级聚类算法。(b) 芯片数据的发散层级聚类算法。20 个基因的数据集是利用 S-PLUS 软件进行聚类。y 轴值表示点之间的不相似度，x 轴代表的是基因（或者样本）的均匀空间，从图中可以看出两个聚类树的整体拓扑结构很相似，但也有几个基因（基因 1、基因 10 和基因 12）在收敛和发散聚类树中位置不相同（箭头所示）

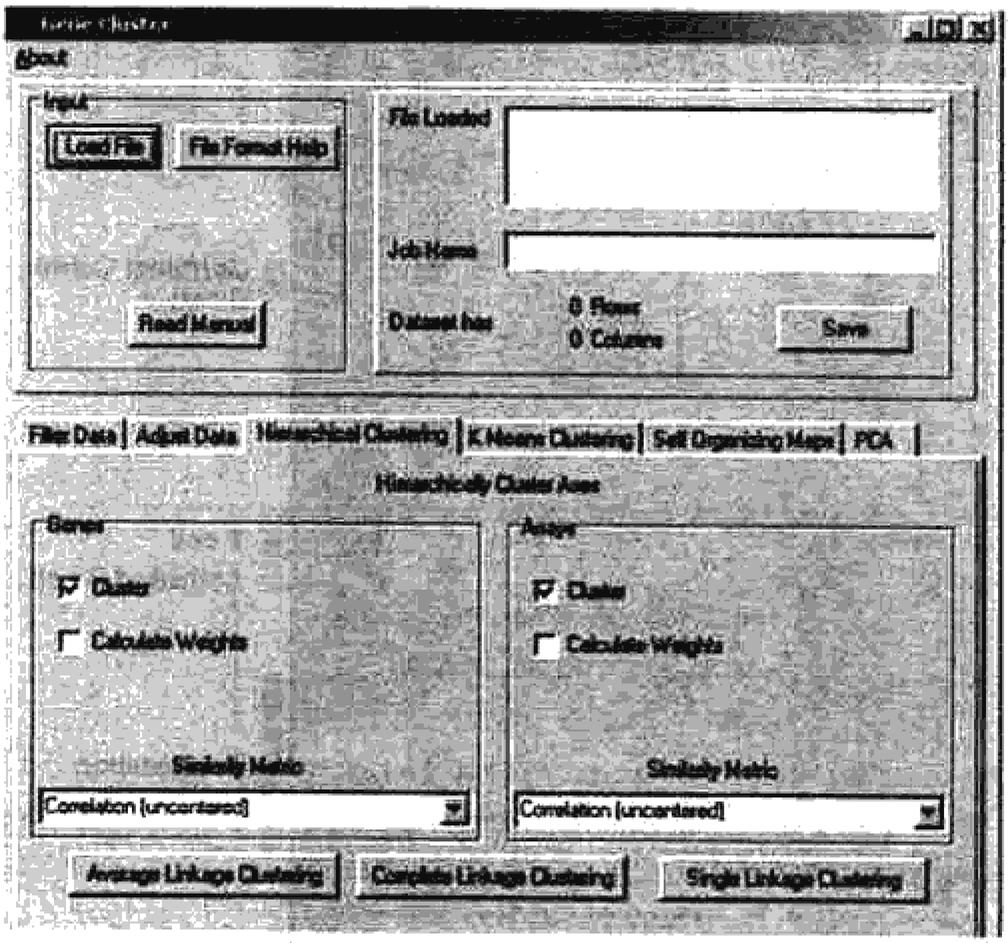
关系数 r，算法为收敛层级聚类。图 7-12 显示了举例的输出结果。基因表达值通过设置不同的颜色表示不同含意：亮红色表示上调基因，亮绿色表示下调基因。这样可以让我们在图上能看到更多的数据，并且更容易明白它们的含意。

在聚类过程中，对基因和芯片样本都进行聚类［双向聚类（two-way clustering）］的目的是用来寻找在所有的样本中基因表达的模式。对于这方面的问题，Alizadeh 等（2000）给出了一个比较好的例子。他们通过基因表达谱给出了恶性淋巴细胞的亚类（图 7-13）。

对于层级聚类还可以举出另外的例子，我们分析了用铅、钠和不用金属处理的星细胞

	A	B	C	D	E	F
1	YORF	0 minutes	30 minutes	1 hour	2 hours	4 hours
2	YAL001C	1	1.3	2.4	5.8	2.4
3	YAL002W	0.9	0.8	0.7	0.5	0.2
4	YAL003W	0.8	2.1	4.2	10.1	10.1
5	YAL005C	1.1	1.3	0.8		0.4
6	YAL010C	1.2	1	1.1	4.5	8.3

(a)



(b)

图 7-11　两个最重要的芯片数据分析软件 Cluster 和 TreeView，由斯坦福大学的 Michael Eisen 在 Patrick Brown 的实验室开发。(a) 数据输入格式需要表格文档（如 Excel），表中，行代表了基因，列代表了样本类型或观察值。(b) 该软件可以对数据进行各种分析，包括层级聚类。Cluster 和 Treeview 可以从网站 http://rana.stanford.edu/software 下载

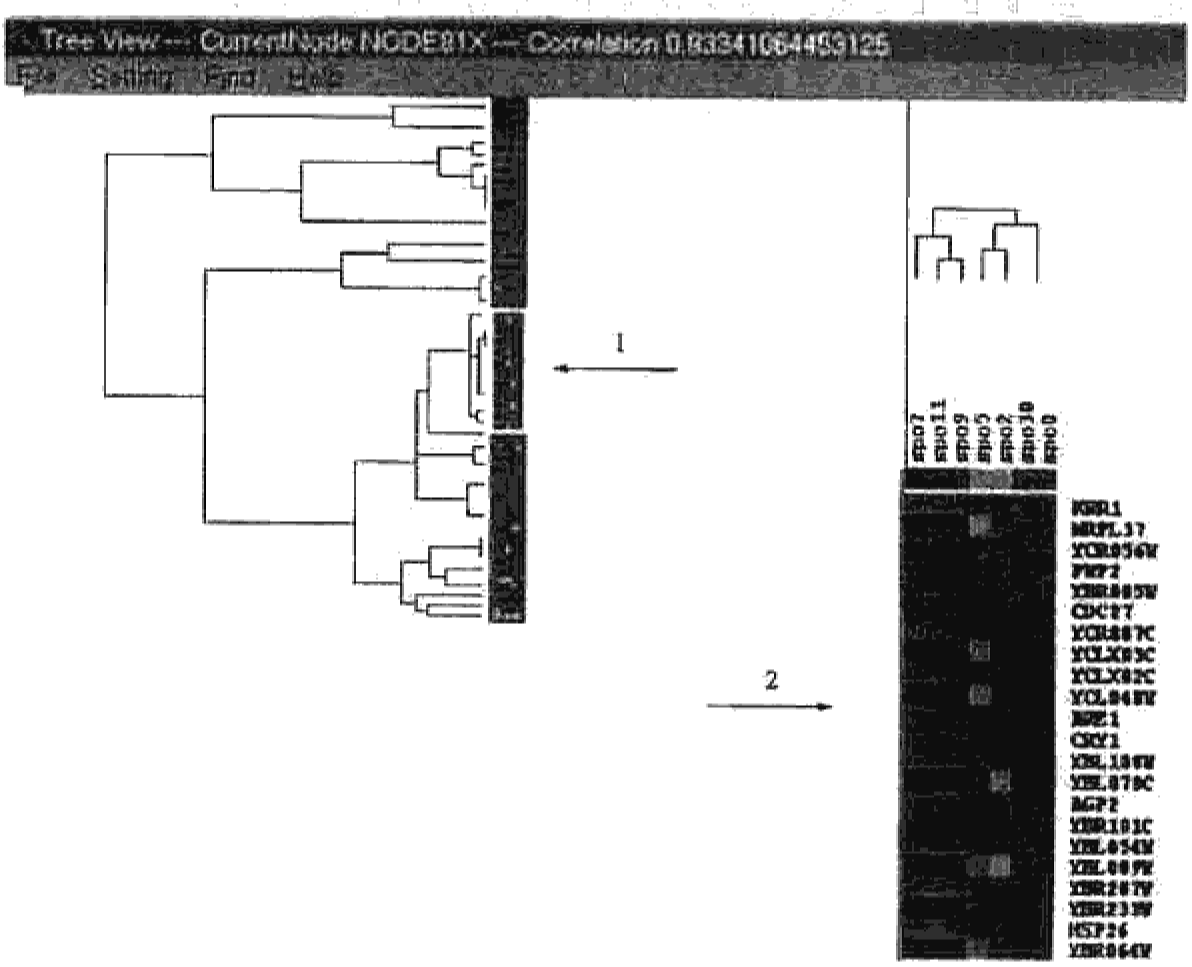


图 7-12　TreeView 软件输出结果举例。输出的结果中，基因被标以不同的颜色，可以对它们上调或下调的含义进行解释。可以通过基因的名字或者在节点之间移动对树进行搜索。例如，通过选择树的一个区域（箭头 1），相应的基因信息就会在右边被放大（箭头 2）

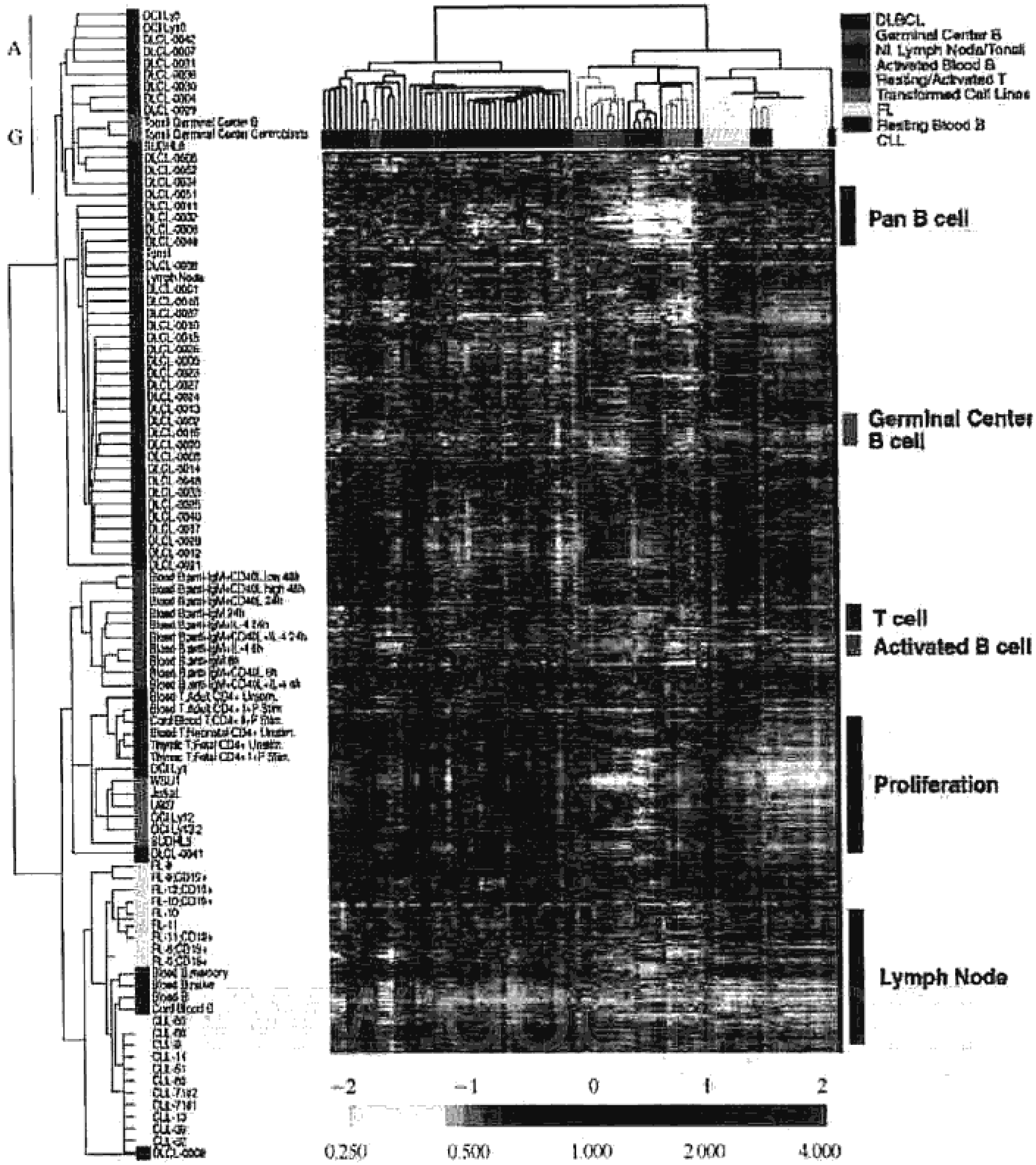


图 7-13 层级聚类的双向聚类举例。Alizadeb 等（2000）对 96 个正常和恶性淋巴细胞样本中 180 万个基因表达值进行聚类分析。细胞系是以列进行聚类的，每一行为基因表达值。分析者用 17856 个 cDNA 克隆制作了芯片。研究表明，癌症病人中有大量 B 淋巴细胞扩散情况的可以通过它们的基因表达谱分离出来。不同类型的癌症病人具有严格的表型，并且这种异质型可以在分子水平上反映出来。这些数据的聚类通过 Cluster 和 TreeView 实现，授权使用

(astrocyte）的基因表达谱。在美国，铅是对环境造成污染的最常见的金属，铅能造成人的智力发育迟缓并且导致人的活动出现问题。通过用铅或者无毒的金属钠处理星细胞，可以得到铅污染对细胞的分子活动影响的情况。这个工作的一个目标是为正常和疾病样本中的基因表达定义分子标签（分子谱）。另外，还可以通过对基因的聚类找到具有相同或者相似生物学功能的基因。在这个实验中，我们处理 11 个细胞系，并且得到的结果可以区分出对照和铅处理样本（图 7-14）（Bouton 等，2001）。对聚类树的进一步分析使我们可以找到哪些单独的基因具有能够区别对照细胞系和铅处理细胞系的能力。这些基因也许能够影响细胞的功能并且最终导致在铅中毒情况下的人活动能力受损或缺失。

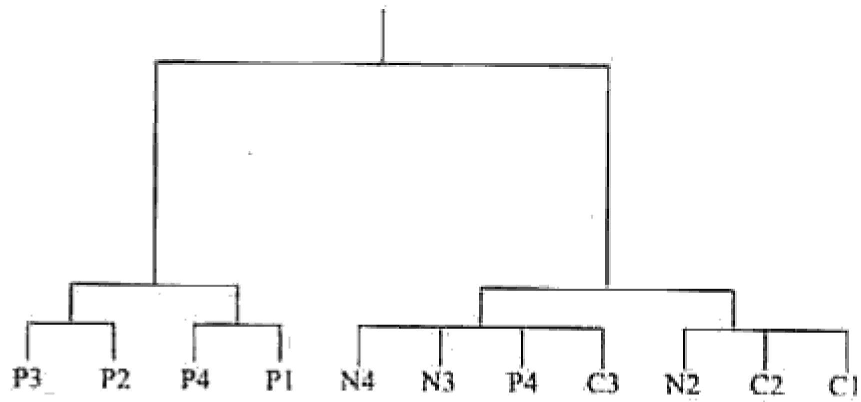


图 7-14 聚类树举例。星细胞被铅处理（P），铅是环境中对儿童健康损伤最大的金属，或被钠（N）处理，或不用金属处理（对照，C）。利用 Clontech 芯片得到 588 个基因的表达值，实验样本被聚类（Bouton 等，2001）。该实验显示通过铅处理的细胞中基因的表达具有标记或谱，并且在铅处理中表达值改变的某些特殊基因也许可以让我们弄明白导致儿童认知和活动不正常的分子机制

7.4.2 分离方法用于聚类：k 均值聚类

在数据聚类分析的具体过程中，有时我们知道数据应该分几类最合适。例如，我在对照实验中只做了一个样本，或者时间序列中选取了 3 个时间点。在非监督聚类算法中，有一类型的算法为分离方法，有此方法可以把数据分为 k 类（Tavazoie 等，1999）。当数据被用户分为 k 类后，每一类中必须至少包含一个对象 n（如基因表达值），并且，每一个对象必须完全是属于它所在的类（在所有的例子中，$k \leqslant n$），任意两个不同类不能用任何一个共有的对象，并且 k 类加和数据集正好是一个完整的数据集。

在 k 均值聚类（k-means clustering）中，k 值如何选取？如果芯片实验中包含两种不同的疾病样本和一个对照样本，你可以把 k 值定为 3。另外，k 值也可以让计算机程序选取，当然预先要设置一些 k 可能的取值。利用 k 均值聚类的输出结果得不到聚类树的结果，因为虽然数据被分离为 k 个类，但这些类之间没有层级结构。

> 许多类型的聚类算法可以用 S-PLUS 统计软件包来实现或者也可以用 Cluster/TreeView（见网络资源，表 7-4）。

k 均值聚类算法是一个迭代算法。在聚类开始，可以随意指定任意对象（如基因）属于某一类，然后，要把某一个类的中心计算出来（可以根据聚类参数定义），另外，类中心的定义就是选取到已知类的中心最远的点为该类的中心，最后，每一个点就被分到距离它最近的类中。在连续迭代过程中，一些对象被重新分派到其他的类中，直到每一个类内距离与类均值差的平方和最小时迭代停止，做了多次迭代后，每一个类内包含的基因都有了相似的表达谱。Tavazoie 等（1999）做了芯片实验，他们用 k 均值聚类的方法发现了酵母中转录调控网络。

7.4.3 聚类策略：自组织图

自组织图算法（self-organizing map，SOM）类似于 k 均值聚类算法。Todd Golub、Eric Land 和 Whitehead 研究所的一些科学家首先利用自组织图算法对芯片数据进行分析（Tamayo 等，1999）。

对于 SOMs 或者其他结果的聚类算法，可以预先规定你所分析的数据可以聚为多少类（例如，根据芯片实验时的条件，如对照、实验样本等），这样在聚类过程中，就可以确定数据初始分多少类。

k 均值聚类算法输出的结果没有没有层级结果，而自组织图算法与 k 均值聚类算法不同，它分离出来的类之间具有一定的结构（图 7-15）。自组织图算法中，首先要选择一定数目的“节点”（与 k 值大小相近）并且对“节点”的几何形状初始化，例如 3×2 矩形的格子（把节点用实线连接起来，构成格子）。在 k 均值聚类过程中，利用临近类之间数据的相互关系，用迭代的方法聚类。在自组织图算法的迭代过程中，节点为了更靠近已知的数据需要向数据点移动，移动的轨道如图 7-15 所示。数据点在图中用黑点表示，SOM 的 6 个节点用大的圆圈表示，移动的轨道用箭头表示。所得的结果就如图 7-12 所示的一个聚类树。

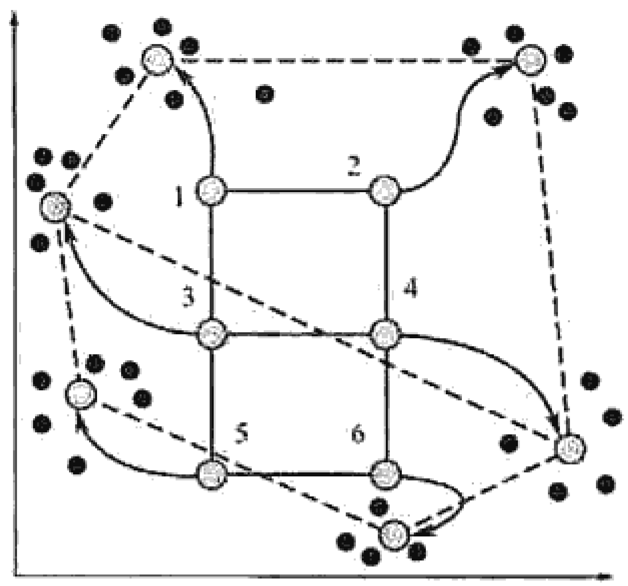


图 7-15 自组织图输出的聚类结果具有一定的结构。它的这种特征与 k 均值聚类不同，k 均值聚类只是把数据归为预定的几个类中，分类后结果没有结构性。自组织图算法首先由 Chu 等（1998）作为非监督的神经网络算法用于寻找酵母孢子生长时的共调控基因

7.4.4 主成分分析算法：芯片数据的观察算法

主成分分析（principal components analysis，PCA）属于记述统计学方法，用来分析芯片实验的基因表达数据从而找到不同类型的基因表达模式。主成分分析算法不但比较容易应用，并且还具有强大功能，能把复杂的数据集转换为比较简洁的数据集。PCA 可以降低数据集的多维空间，从而能将数据集在 2 维图或 3 维图上表现出来，以至于能从中找到类中对象（即实验的基因或样本类型）之间的相互关系。PCA 已经被用来分析酵母和哺乳动物系统中的基因表达值（Landgrebe 等，2002；Misra 等，2002；Alter 等，2000；Wall 等，2001；Bouton 等，20001）。

PCA 也叫单值分解（singular-value decomposition）（Alter 等，2001）。它是一个线性工程方法，也就是说，在分析数据时，需要把数据转换到低维空间上。PCA 的相关工程方法包括独立成分分析（independent components analysis）、因子分析（factor analysis）、多维尺度分析（multidimensional scaling）以及相关的一些分析。

PCA 的中心思想是把比较大向量转换为比较小的相互之间没有关系的向量（称为主成分）。用 PCA 操作的向量可以是许多基因的表达值（如 10000 个基因的表达值），或者不同样本中的基因表达结果，或者既包含基因表达值又包含样本类型。在一个标准的芯片实验中，PCA 的点可以被探测到并且它还去除一些冗余的点（如一些在对照和实验样本中表达量没有改变的基因）。去除冗余点的目的是为了降低数据集的噪声，并且还可以鉴别出一些孤立的点（或者孤立的点构成的类），而这些点往往是我们研究所感兴趣的。

我们还是以图 7-8(a) 中的数据矩阵为例，该矩阵包含 20 个基因表达值和 3 个样本（时

间点)。我们可以用 PCA 把这个矩阵转换到图 7-16(a) 上。在该图上，20 个基因用 1～20 的数字表示。利用主成分分析绘制的图与图 7-10 聚类树相比，在每个例子中，我们都可以找到基因的不同类（例如 5 和 17 或者 4、9 和 15）。

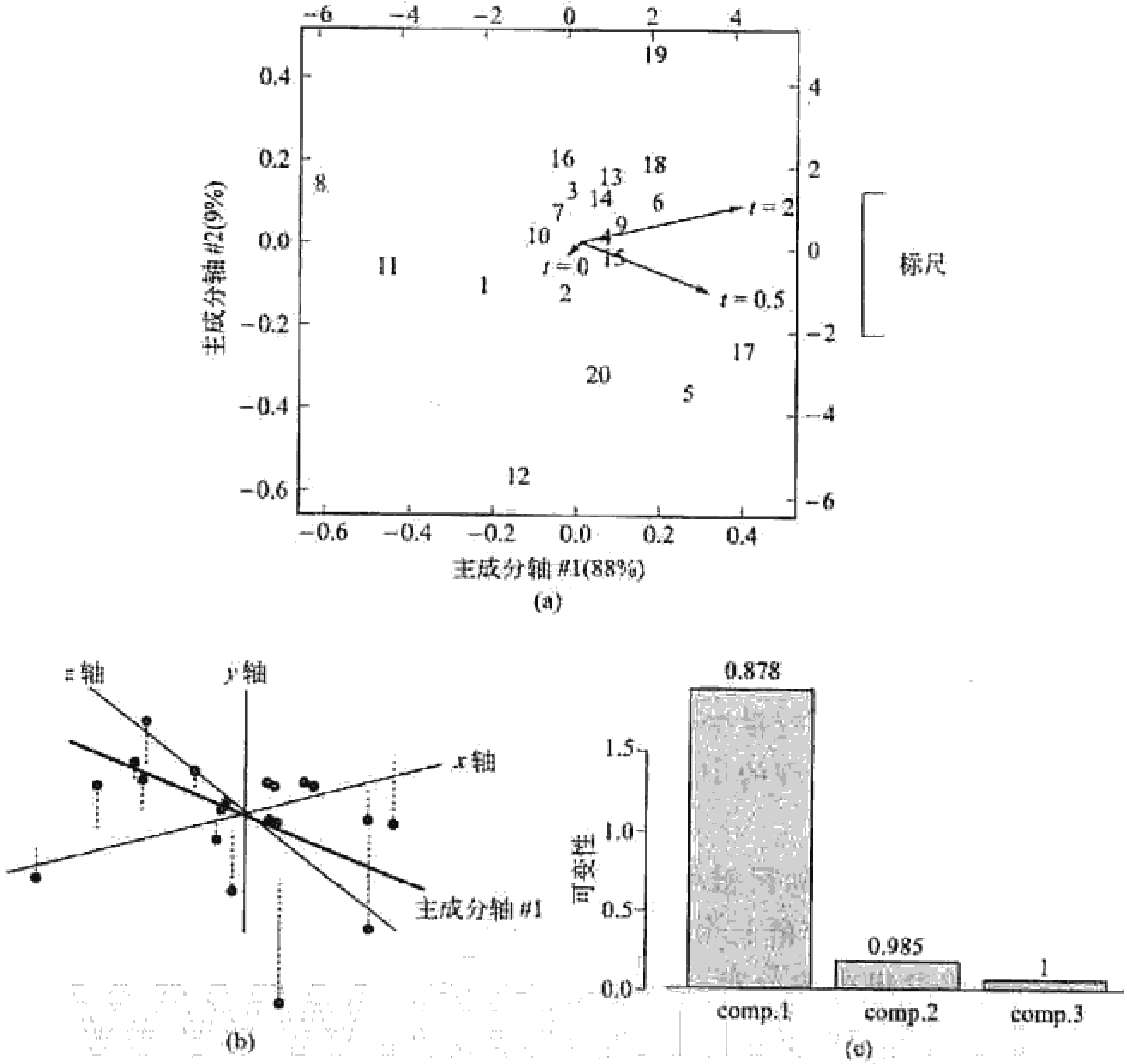


图 7-16 PCA 分析降低了芯片数据的向量使基因或者样本之间的相关性可以显示出来。(a) 显示了 PCA 分析绘图。第一个和第二个主成分轴在图上显示出来，并且可以显示它们占用数据向量的百分比。含有 20 个基因表达的数据集［图 7-8(a)］被绘制到图上，在向量空间中，距离比较近的基因具有更相似的基因表达谱。含有 3 个变量的生物标尺应用到 PCA 的作图分析中，它可以显示 3 个时间样本（$t=0, 0.5, 2$）的相对分布情况。第一个主成分轴是在 (b) 中通过数据点的中心（原点）并且最能拟合所有数据的直线，它代表了数据集变量的大部分信息。第二个主成分（没有显示）仍然要通过图上的原点，并且要和第一个主成分轴垂直。(c) 是一个柱状图，它显示了每个主成分轴占有的所有原始数据变量信息的百分比（见每个柱上面的数字）。在图中，可以看到第一个主成分轴占据了数据信息的绝大部分（88%），这也提示在看 (a) 时，x 轴应该被足够重视。例如，在 (a) 中，根据第一个主成分轴判断，基因 8 和基因 11 远离其他变量，它们肯定受到了比较重要的调控

图 7-16 是用 S-PLUS 软件（insightful 公司）绘制的。数十个不同的软件都可以运行 PCA，其中一些在表 7-4 中列举出来了，并且可以从列举的网络资源中找到。

在 PCA 运行过程中，首个主成分应该尽可能体现数据的可变性变量。第二个主成分就要比第一个主成分的重要性小一些，并且每个主成分轴相互之间必须为直角，这就意味着它

们相互之间没有任何关联（见下文）。PCA分析中，最基本的是二维作图的方法，首个主成分占据x轴，且第二个主成分占据y轴［图7-16(a)］。但是，三维作图在PCA分析中也比较常用。如果再有其他的主成分轴，因为它们仅稍微体现了数据矩阵的可变性，因此一般的作图也不会显示出来。

PCA分析都是从m个实验值（基因表达值）和n个变量（实验条件）的矩阵开始，分析的目的是为了降低数据矩阵的向量，从而使数据矩阵的向量到一个新值$r(r<n)$。新的r个向量会尽可能体现出原始数据矩阵的可变性。PCA分析的第一步是创建一个新的$n\times n$的矩阵。这个矩阵可能是协方差矩阵或相关系数矩阵［在我们的例子里（如图7-16），它是一个3×3的协方差矩阵］。选取主成分（特征向量）作为最大的变量（特征值），这样处理的实际意义就是，在我们的样本数据集里，如果某个基因在3个时间点上的表达量没有变化或者变化很小，那么它将不会被包含在主成分中。

那么，在主成分分析中，第一个主成分轴与原始数据有怎样的相关性呢？把原始数据在三维空间图上表示出来，并且重新调整x轴、y轴、z轴，使原心（质点）落在所有数据点的中心［图7-16(b)］。寻找最适合这些数据点的直线，这个直线就为第一个主成分的轴。调整该直线，它就为图7-16(a)中的x轴。第二个主成分轴仍然要通过图7-16(b)的原点，并且必须要和第一个轴垂直。只有这样，它们才不会有相关性，每个轴占有的数据集的可变性越来越少。图7-16(c)的柱状图显示，第一个轴占有了数据88%的变量，而第二个和第三个主成分轴比第一个占有的变量少得多（只有9%和3%）。因此，第一个主成分体现的数据变量大约是第二个主成分体现的10倍。在图7-16(a)上，第三个主成分轴没有显示，因为它体现的数据变量很小。

在图7-16(a)中，绘图标尺被引入用于PCA的分析。该标尺是个向量标尺，通过它可以在图上画出用PCA分析三个时间点（$t=0$、0.5、2）得到的相对分布。对于$t=0$的数据，它只有相对很小的分布，这对观察原始基因表达数据比值［图7-8(a)］也是很有意义的。

PCA分析的最后一步就是作二维或者三维图，图上包含的点表示基因或样本类型。如果用PCA对样本类型作图，那么，如果任意两点相邻比较近，就表示这两个样本的总体基因表达模式是相似的；相反地，如果这两个点相距比较远，则表示这两个样本总体上具有不同的基因表达谱。PCA分析中，初始数据集是高维的，对于含有1000个基因表达值的数据，理论上可以说它有1000维的空间。PCA使原始数据集的高维空间降低到二维或者三维。在降低数据的维数时，PCA的目标是尽可能保留原始数据集的信息。

Raychaudhuri和他的合作者（2000）用PCA的方法分析了Chu等（1998）的整个数据集，该数据集包含7个时间点的6118个基因表达值。他们的PCA结果显示，只需两个主成分就可以包含原始数据总可变量90%以上的信息。他们进一步的处理是把主成分（1）作为基因改变水平上的信息归纳，主成分（2）作为整个时间改变水平上的信息归纳。总的来说，主成分轴可能可以，也可能不可以对应于有明显生物学含义的变量。这主要是因为主成分的特征（即尽可能多地占用数据集的信息），而特征的判断严格依据变量的标准。

> 在PCA分析中，向量空间大小可以和原始数据矩阵的样本大小一样大，不过，利用PCA得到的信息主要包含在PCA的前2个或者前3个主成分中。

在第8章对蛋白质家族数据库Pfam的分析中，还会再次遇到PCA。PCA基于蛋白质序列比对矩阵来分析蛋白质之间的相关性，同时，PCA也用来探究

在COG（clusters of orthologous genes）数据库中（第14章）不同基因组之间的相关性。

7.4.5 监督算法对芯片实验中基因或样本数据的分类

在芯片数据分析中，基因表达值之间的距离和相关性可用两种方法来分析：监督算法和非监督算法（图7-1）。对非监督算法已经讨论了很多，并且知道利用非监督算法可以找到大数据集的某些模式。监督算法分析不同于非监督算法分析，因为用监督算法分析时，实验者已经知道了这个实验的一些基因或样本的信息。例如，在细胞系或者正常和肿瘤样本中（例如，Alizadeh 等，2000；Shipp 等，2002；Khan 等，1998；Perou 等，1999；West 等，2001）转录型（transcriptional profiling）已经表现出来。在有些实验中，肿瘤样本还会被进一步细化（如良性肿瘤和恶性肿瘤等）。而这些例子也可以用非监督算法来分析。

监督算法用于芯片数据的分析目标是制定一个能够把基因或样本进行分类的规则。在每个例子中，算法都是从已知类（如正常对肿瘤样本）的基因表达值入手，对这些样本的数据进行训练使它们能够得到这些类的分类规律。正和负的样本数据都得用来进行训练，最终得到分类规则。然后，再用这种规则对未知类的样本进行分类，并且对这种算法得到的预测或者分类精度进行评估。

监督的数据分析算法中最常用的有支持向量机（support vector machine）、监督的机器学习算法（supervised machine learning）、神经网络（neural network）和线性判别分析（linear discriminant analysis）等。下面举一个监督算法的例子，Brown 等（2000）利用支持向量机把酵母基因分为六个功能类：三羧酸循环（tricarboxylic acid cycle）、呼吸作用（respiration）、细胞质核糖体（cytoplasmic ribosomes）、蛋白酶（proteasomes）、组蛋白（histones）和螺旋-转角-螺旋蛋白（helix-turn-helix proteins）。他们用了3次交叉验证：数据集被分为3个部分（数据集1，数据集2和数据集3），首先，数据集1和数据集2用来训练支持向量机，然后用该算法测试“未知”的数据集3；其次，数据集1和数据集3用来训练，数据集2用来测试；最后，数据集2和数据集3用来训练，数据集1用来测试。所有这些做完后，他们得到了算法的假阳性率并且得出支持向量机算法比非监督聚类算法和部分的监督聚类算法优越。

7.4.6 芯片数据注释

DRAGON 可以从网站 http://pevsnerlab.kennedykrieger.org 上得到。相关的工具（如用来进行基因组研究的RESOURCERER）可以从网站 http://pga.tigr.org/tigr-scripts/magic/r1.pl 上得到。斯坦福大学关于 Clones 和 ESTs 的在线网络资源的网站为 http://source.stanford.edu。ARROGANT 软件的网址为：http://lethargy.swmed.edu。

芯片实验的主要任务就是弄清得到的基因表达模式的生物学重要性。研究者经常依赖于手工检索的文献知识和专家的经验来解释芯片的结果。许多软件工具接受增加数字（对应于芯片实验中的基因）的列表并提供注释。例如，在DRAGON数据库（database referencing of array genes online）的网页上允许芯片数据用从公共数据库［如UniGene，图7-17(a)］（Bouton 和 Pevsner，2000；Bouton 等，2003）中得到的数据来注释。DRAGON 提供一些可视化工具让用户可以鉴别出发生在基因或者蛋白质家族中（Bouton 和 Pevsner，2002）基因表达值的改变情况。某些注释工具（如DRAGON）的目标是为寻找基因表达的重要生物学意义提供知识。

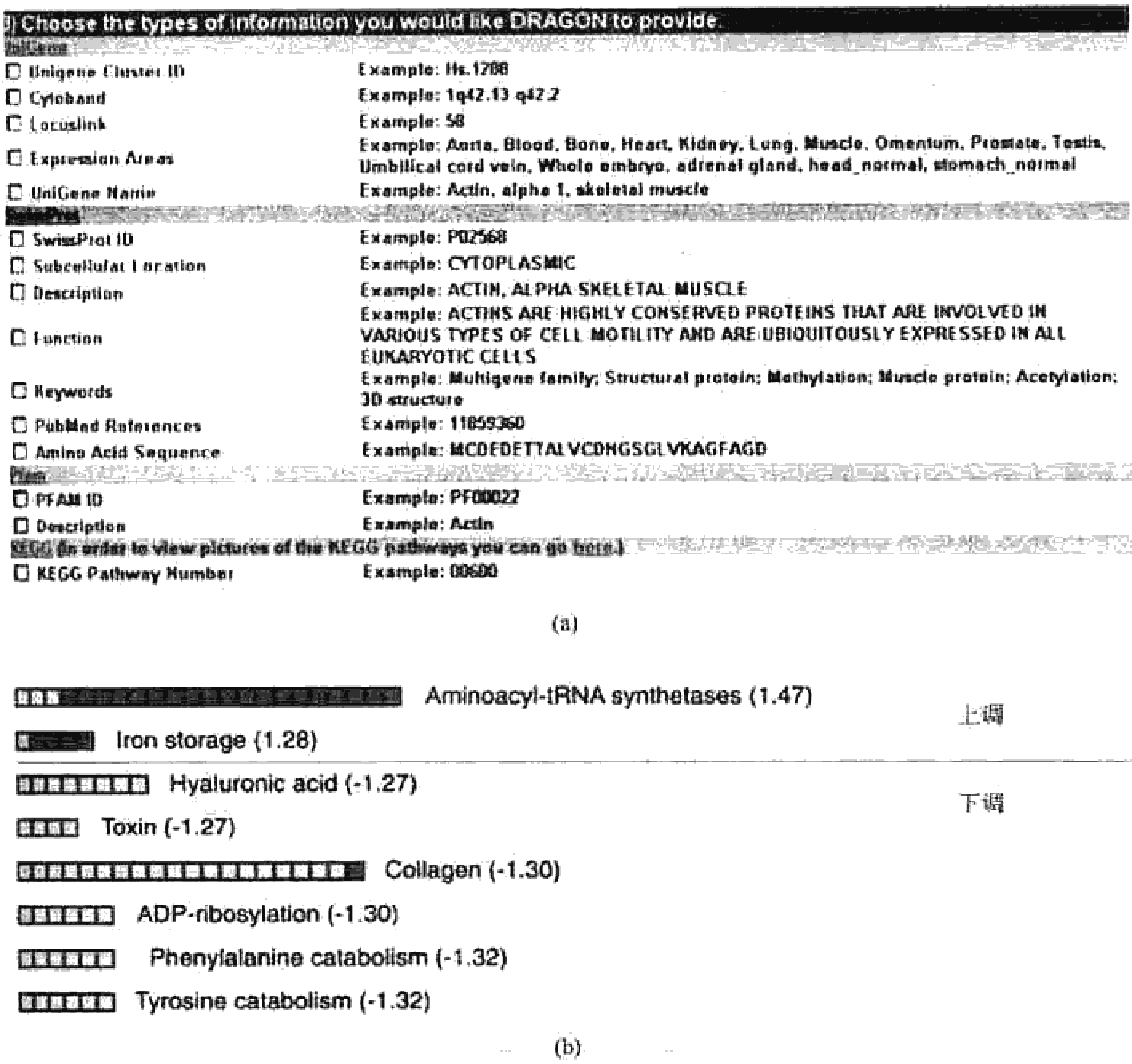


图 7-17 DRAGON数据库允许库内的芯片数据被注释和数据的可视化。(a) 可以让用户输入芯片数据和注释，数据的格式是必须是规定的格式。图中包含对应每个基因标签的数据，代表的意义有基因在染色体上的定位或者基因翻译得到的蛋白的功能等。这些数据主要来自如下数据库：UniGene、SwissProt、Pfam和KEGG（见第8章和第10章）。(b) 铅处理的细胞中基因的表达情况。编码氨酰基(aminoacyl) tRNA合成的基因上调，其他家族的基因有的上调，但多数都是下调。图中不同的颜色代表基因受到调控的程度，并有对LocusLink的超级链接。引自Bouton等(2001)

7.5 展望

DNA芯片技术可以让实验者快速和高通量地测出在一个生物学样本中成千上万的基因表达值。这种技术在20世纪90年代后期变成了研究多种生物学问题的工具。在芯片实验中，几千甚至几百万的数据点都能够测出。这样，芯片数据的分析问题就必须利用一些数学工具，这些数学工具都是在分析其他包含多数据分析的科学分支的问题上发展而来的。这些工具包括聚类分析(clustering analysis)、主成分分析(principal-components analysis)和其他一些把高维数据集降为低维的可用格式数据集的方法。对于芯片数据分析来说，主要面临的需要解答的问题如下。

① 比较两种实验条件（例如，被处理过的细胞系和没有用药的细胞系）下，哪些基因受到了引人注目和重要的调控？

② 比较多种实验条件（例如，对正常和染病个体中100个细胞系基因表达的分析）下，

哪些基因始终受到比较重要的调控？

③ 对于芯片数据，是否可以通过聚类的方法把样本或者基因按功能分类？

而对于芯片实验的数据分析，最大的挑战就是能够让实验数据处理得到的类（模式）等结果用生物学的机制来解释。

最后要说的是，DNA芯片现在不仅用于测定生物样本中的基因表达量，同时还有其他很多方面的应用。它们的主要应用有：探测多态现象（见第17章），直接测定基因序列，确定调控的DNA序列和在DNA中的插入和复制情况（见第18章）。芯片实验的应用范围在近些年还会扩大。

7.6 常见问题

在芯片实验的不同阶段会出现很多问题，这些问题包括以下一些方面。

① 芯片实验设计需要严谨的过程，不过在芯片实验中还是有许多问题没有注意或者解决。多少数量的对照和实验样本才是最适合的是芯片实验设计的一个基本问题。另外，最适合的实验重复次数是多少？这也是一个很重要的芯片实验设计问题，而现在对于重复次数的问题还没有一个统一的标准。

② 芯片实验中，基因的表达值是用颜色亮度值来确定的，而颜色亮度值和细胞中mRNA的精确数量的相关性到底有多少还很难确定。而这种情况随着芯片实验每一步误差的影响还会进一步恶化。这些过程有：从把RNA提取出来并转化为带有射线或者荧光的探针、杂交的效率到数据分析的多种误差。现在有些研究小组提出了RNA分子标准集的全局化标准的分析方法，不过这些方法还没有被广泛地接受。

③ 在数据分析时，需要注意全局和局部背景噪声信号的校正。

④ 在探测分析（记述统计学方法分析）时，需要引入适当的参数量（例如皮尔逊相关系数），利用这样的参数，可以确定受到调控的基因。

⑤ 每个芯片数据的分析方法都有其自身的优点和局限性。例如，比较常用的非监督算法（如聚类分析）会丢掉被研究的样本本来所属类（如从癌症病人不同亚型中选取的细胞系等）的信息。而监督算法对于类的假设可能是错的。

⑥ 许多实验的人为误差可以通过精确的数据分析发现。在散点图中，相位差的问题表明生物样本受到了污染。聚类分析可能暴露相容性的困难，这不是因为实验和对照之间的差异，而是因为样本操作者的影响或不同时间的取样造成的。

7.7 网络资源

相关的网络资源见表7-3和表7-4。

表7-3 芯片数据分析资源

资源	评注	网址
Y. F. Leung's芯片资源	芯片资源链接列表	http://ihome.cuhk.edu.hk/~b400559/arraysoft.html
斯坦福芯片数据库	芯片数据处理软件工具列表	http://www.dnachip.org/index.shtml
美国基因组研究所	各种程序	http://www.tigr.org/software/
芯片和数据分析	芯片分析的文献	http://linkage.rockefeller.edu/wli/microarray/index.html
mAdb	为美国癌症研究中心提供信息技术	http://nciarray.nci.nih.gov/

表 7-4 芯片数据分析和芯片数据分析统计软件包（其他几十个可以得到的软件包）

软件	公司/组织	网址
Array Suite	Affymetrix	http://www.affymetrix.com
Array Organizing Tool (ARROGANT)	University of Texas Southwestern Medical Center at Dallas	http://lethargy.swmed.edu/
Atlas Navigator	Clontech	http://www.clontech.com
BioArray Software Environment(BASE)	Lund University	http://base.thep.lu.se/
BioMine	Gene Network Sciences	http://www.gnsbiotech.com/products.shtml
Cleaver	Stanford University	http://classify.stanford.edu/
Cluster Analysis of Gene Expression Dynamics(CAGED)	Harvard Medical School	http://www.genomethods.org/caged
Cluster/Treeview	Stanford University	http://rana.lbl.gov/
DRAGON	Kennedy Krieger Institute	http://pevsnerlab.kennedykrieger.org
Expression Profiler	European Bioinformatics Institute	http://ep.ebi.ac.uk/
GeneCluster	Whitehead Institute	http://www-genome.wi.mit.edu/cancer/software/software.html
GeneCruiser	Whitehead Institute	http://www-genome.wi.mit.edu/cancer/software/software.html
GeneSight	BioDiscovery	http://www.biodiscovery.com
GeneSpring	Silicon Genetics	http://www.sigenetics.com
GeneTraffic	Iobion	http://www.iobion.com/
GeneLinker	Molecular Mining	http://www.molecularmining.com
J-Express	MolMine	http://www.molmine.com/
Partek	Partek	http://www.partek.com
Pathways	Research Genetics	http://www.resgen.com
S-PLUS	Insightful	
Signficance analysis of microarrays(SAM)	Excel plug-in	http://www-stat.stanford.edu/~tibs/SAM
SNOMAD	Kennedy Krieger Institute	http://pevsnerlab.kennedykrieger.org
Spotfire Decision Site	Spotfire	http://www.spotfire.com/
SPSS-Clementine	SPSS Inc.	http://www.spss.com/spssbi/clementine/
TIGR MEV	The Institute for Genomic Research	http://www.tigr.org/software/m
X-MINER	X-Mine	http://www.x-mine.com/
XLSTAT	Excel plug-in	http://www.xlstat.com

7.8 问题讨论

1. 芯片数据可以通过多种方法聚类，并且得到多种不同的结果。而怎样才能知道哪个聚类结果是“正确的”（与生物学意义最相关的）？

2. 确定基因是否受到重要性调控的最好的标准是什么？如果你引用倍数改变作为标准，也许会有这样的情况：某种倍数改变在统计学上有重要性与在生物学上是否有重要性也许并不统一，如果通过校正并发现在芯片实验中没有发生基因表达水平的很大改变，那么，是不是说生物学问题就是一个比较复杂的似是而非的问题。

7.9 问题

1. 访问和下载芯片数据。进入斯坦福大学的芯片数据库 http://www.dnachip.org。

① 点击公共数据，选择 Chu 等（1998）发芽酵母孢子形成时的转录系统数据库（Science 282：699-705）。

② 选择“Spo 电子数据表”。并且把它保存在电脑的硬盘里。

③ 打开 Excel 软件，并利用该软件打开刚才保存的文档文件。

④ 列 A 包含了在芯片实验中每个基因的名字，列 B 和列 D 是得到的在时间 $t=0$ 时样本所染绿色荧光和红色荧光的值。列 C 和列 E 是芯片实验中格点的背景颜色值。

2. 在不校正背景荧光值的情况下创建 Excel 表。在此分析部分，我们将忽略背景荧光值（列 C 和列 E），并且仅仅处理时间 $t=0$ 时格点的颜色值。对原始数据进行以 10 为底的对数转换。见本章前文讨论的关于原始数据对数转换后的优点分析。

3. 对原始数据作图。

① 点击列 B 上面的 B 字母，选择整列。按住键“Ctrl”，然后选择列 D，同样按列 D 最上的字母 D。这样就可以在选择 B 的同时也选择了 D（并且，列 B 和列 D 之间的列 C 没有被选择）。

② 选择 Excel 页上的插入选项，然后点击制图，选择 XY（散点图）并且点击“下一步”，图表出现，该图上 y 轴的值代表是红颜色值，x 轴的值代表绿颜色值。该图与图 7-3(b) 类似。

③ 选择“下一步”。这一步让你做所得表的标题和轴的标题。

④ 继续选择“下一步”。在此你可以选择“作为新工作表插入”或“作为其中的对象插入”保存，如果选择“作为新工作表插入”，该表会在 Excel 表上显示出来。

⑤ 选择新的表或对象后选择“完成”。

附记：当创建图时，你选择的第一列数值就为作为 x 轴的值且第二列数值作为 y 轴的值。

4. 做对数转换后的数值图。

① 在列 B 后面插入一新列。插入新列的方法为选择列 C 上面的字母“C”，然后点击右键，选择“插入”。在这个新列里，你可以把列 B 中的数值转换为以 10 为底的对数（“t0 green”）。

② 点击第二列（即新的 C 列），然后移动鼠标，点击列上面一行空白栏左面的“=”号，然后在空白栏中会出现“=”号，在“=”后面输入 lg(B2)，按“Enter”键，在 C 列的 C2 就会出现数字 3.552668。

③ 现在，对 C 列其他空白的地方复制上一步的操作。具体过程为：点击 C2 栏，然后选择 Excel 的“编辑”菜单，再选择“复制”项，然后选择 C 列中的所有空白栏，再选择 Excel 的“编辑”菜单，再选择“粘贴”项。这样，在 C 列中的数字就为对应的 B 列数字的以 10 为底的对数值。

④ 对“t0 red”对应的列重复以上各步操作。

⑤ 对对数列的值做散点图，如图 7-3(c) 所示。

5. 在考虑到背景信号噪声的情况下创建 Excel 图。在这一步分析中，要结合背景荧光

值的信息［在原始的 Excel 表中，为 C 列和 E 列，如果考虑到先前插入的两列（4. 步骤），那么现在的表中背景荧光值的信息在 D 列和 G 列］。“t0 green bkg”列和“t0 red bkg”列的值是芯片上每个格点附近空白处测得的值，它们说明了非特异结合固定 RNA 能力的大小。我们将把每个格点的荧光值去除上面测得的非特异结合能力的值，因为在芯片上，不同格点处的非特异结合能力会不同，或者不同的绿和红的水平，非特异结合的能力也不同，因此要去除这个背景噪声。

① 在 D 列（“t0 green bkg”）后面插入一个新列。利用列上面横的空白栏左侧的“＝”号，点击它，然后在空白栏的“＝”后面输入“B2-D2”（即“t0 green-t0 green bkg”），这样就把“t0 green”的背景噪声去除了，操作的过程中，要确保实验的数据都是原始数据，而不是经过转换的对数数据。对“t0 red bkg”和“t0 red”重复上面的操作，去除噪声。

② 创建新的两列，且计算上面去除背景噪声后所得数值的对数。

③ 对新的两列以 10 为底的对数值作图。如果在芯片实验中，红和绿的原始数据的背景噪声都一样，那么新图和没有做去噪处理的图之间相差不会太明显［见图 7-3(c)］。

6. 对所有的数据归一化。因为实验所用的样本之间会有差别，有的样本中会比其他的样本有更多的 RNA，所以把样本中荧光颜色值转换为平均荧光颜色值是很重要的问题。

① 在去噪处理后得到的原始格式数据的后面各插入一列（注：归一化的是去噪处理后原始格式的数据，而不是经过对数转换后的数据）。

② 计算每一列的平均值。具体过程为：点击所处理列最后一个数值的下一栏（如 E6120 栏），点击“＝”号，再在空白栏的“＝”后面输入 AVERAGE（E2：E6119），然后点击“Enter”，在栏 E6120 里就会出现整个列的平均值。接下来再用这一列的每一个值去除以这个平均值。具体过程为：在处理数据的列中，点击第二栏，然后点击“＝”，在空白栏的“＝”后输入 E2/xxxx（这里“xxxx”代表所计算的 E 列的平均值，即栏 E6120 的值），再复制所得的数据，粘贴到其他空白栏中。对红色数据重复以上操作。

③ 另外再重新创建两列，对以上归一化后的数据取对数，然后把所得的两列对数值作图。

7. 对颜色比值的对数值和颜色值的对数均值作图。从这个图中可以观察出颜色比值可作为颜色值对数均值的一个功能，因为基因在低水平表达的变化比基因在转录水平有更大的可变性。从这个图像可以找到数据点所形成图像的曲率。

① 在电子表格上创建三个新列。在一个列中，计算你在 6. 中创建的最后两列每个基因表达的平均数。这些数据代表了每个基因的原始数据去噪归一化后再经对数转换后的红-绿均值。这一列的数据将作为图 x 轴上的数据。

② 在另外一列中，用每个基因原始数据去噪归一化后的红颜色值除以绿颜色值［不能用经过对数转换后的值（如在 6. 中创建的最后一列值），而要用在对数转换之前的值］。

③ 在第三列中，把上一步刚处理得到的数据转换为对数。这样，这一列的值就为原始数据去噪归一化后比值的对数值。这一列的数据将作为图 y 轴上的数据。

④ 作图，所得图将和图 7-4 和图 7-5 中某个图相似。

7.10 自测题

1. 在芯片数据处理中，对原始数据的归一化是必须的，这是因为

(a) 基因表达值不是正态分布；

(b) 芯片实验中，所用的cDNA有的带有荧光，而有的含有射线；

(c) 在不同的样本中，利用荧光或者射线的效率不同；

(d) 不同的样本中，利用荧光的效率不同

2. 芯片数据分析可以利用散点图，从散点图上可以得到的信息不包含

(a) 可以确定哪个基因在相对高或低的水平表达；

(b) 可以确定哪个基因上调表达或下调表达；

(c) 可以确定某个基因和另外一些基因可以形成一类；

(d) 可以确定某个基因属于实验中受到最重要调控基因的前5%

3. 在芯片实验中，基因表达比值的对数转换比原始数据更常用，是因为

(a) 两倍上调基因和两倍下调基因比值的对数有相同的绝对值；

(b) 两倍上调基因和两倍下调基因比值的对数有相同的相对值；

(c) 比值对数的范围和原始数据的范围相关；

(d) 比值对数的散点图是表达值比值散点图的压缩版本

4. 推测统计学分析芯片数据集的假设检验前提为

(a) 概率为何值时就可以表明在比较两个样本后，某个基因受到了重要的调控；

(b) 概率为何值时就可以表明在比较两个或更多样本后，某个基因受到了重要的调控；

(c) 对芯片数据的聚类；

(d) 利用非监督算法或者监督算法分析数据

5. 下面的描述哪个是错误的？

(a) 芯片数据聚类得到的树类似于系统发育树；

(b) 聚类的方法可以应用于芯片的基因或者样本；

(c) 芯片数据聚类可以应用分割算法（例如 k 均值算法）或利用层级聚类算法（例如收敛聚类或者发散聚类）；

(d) 芯片数据聚类算法总是应用主成分分析算法

6. 聚类方法需要应用聚类参数的目的：

(a) 描述哪个聚类树是发散或者收敛的；

(b) 降低高维数据集的维数；

(c) 在含有基因表达值和样本的矩阵中，得到基因表达值的绝对值；

(d) 在含有基因表达值和样本的矩阵中，界定基因表达值之间的关系

7. 自组织图：

(a) 显示聚类后得到结果的结构；

(b) 没有结构的，就如 k 均值聚类的结果；

(c) 含有邻居节点，用此来表示类之间的不同；

(d) 不能用聚类树的方法表示

8. 主成分分析（PCA）：

(a) 熵最小化，并且得到基因和蛋白之间的相关性；

(b) 可以对芯片实验中基因表达数据进行分析，但不能用于蛋白质芯片；

(c) 包含收敛算法和发散算法；

(d) 降低高维向量的维数从而发现基因或者样本之间的关系

9. 芯片数据分析中，非监督算法和监督算法最大的不同在于

(a) 监督算法需要先验知识（如基因或者样本的功能等），而非监督算法不需要；

(b) 监督算法需要指定聚类的数目，而非监督算法不需要；

(c) 监督算法可以对基因或者样本聚类，而非监督算法只能对基因聚类；

(d) 监督算法包含的算法有支持向量机、决策树等，而非监督算法只有聚类算法

7.11 推荐读物

对于芯片数据的聚类分析而言，Michael 和他的合作者（1998）做了人类 8600 个基因作为时间功能的聚类分析。他们的这篇文章非常著名，文章中有对基因表达值相关性参数非常精彩的定义，并且讨论了根据基因表达值的相关性来定义功能类的过程。

参考文献

Alizadeh, A. A., et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511 (2000).

Alizadeh, A. A., Ross, D. T., Perou, C. M., and van de Rijn, M. Towards a novel classification of human malignancies based on gene expression patterns. *J. Pathol.* **195**, 41-52 (2001).

Alter, O., Brown, P. O., and Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **97**, 10101-10106 (2000).

Ball, C. A., et al. An open letter to the scientific journals. *Bioinformatics* **18**, 1409 (2002a).

Ball, C. A., et al. Standards for microarray data. *Science* **298**, 539 (2002b).

Beissbarth, T., et al. Processing and quality control of DNA array hybridization data. *Bioinformatics* **16**, 1014-1022 (2000).

Bouton, C. M., Hossain, M. A., Frelin, L. P., Laterra, J., and Pevsner, J. Microarray analysis of differential gene expression in lead-exposed astrocytes. *Toxicol. Appl. Pharmacol.* **176**, 34-53 (2001).

Bouton, C. M., and Pevsner, J. DRAGON View: Information Visualization for Annotated Microarray Data. *Bioinformatics* **18**, 323-324 (2002).

Bouton, C. M., and Pevsner, J. DRAGON: Database Referencing of Array Genes Online. *Bioinformatics* **16**, 1038-1039 (2000).

Bouton, C. M., Henry, G., Colantuoni, C., and Pevsner, J. DRAGON and DRAGON view: methods for the annotation, analysis, and visualization of large-scale gene expression data. In: G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger (eds.), *The Analysis of Gene Expression Data: Methods and Software*, Springer, New York, 2003, pp. 185-209.

Brazma, A. On the importance of standardisation in life sciences. *Bioinformatics* **17**, 113-114 (2001).

Brazma, A., Robinson, A., Cameron, G., and Ashburner, M. One-stop shop for microarray data. *Nature* **403**, 699-700 (2000).

Brazma, A., et al. Minimum information about a microarray experiment (MIAME)–toward standards for microarray data. *Nat. Genet.* **29**, 365-371 (2001).

Brazma, A., and Vilo, J. Gene expression data analysis. *FEBS Lett.* **480**, 17-24 (2000).

Brown, M. P., et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* **97**, 262-267 (2000).

Chu, S., et al. The transcriptional program of sporulation in budding yeast. *Science* **282**, 699-705 (1998).

Claverie, J. M. Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.* **8**, 1821-1832 (1999).

Colantuoni, C., Henry, G., Zeger, S., and Pevsner, J. SNOMAD (Standardization and NOrmalization of MicroArray Data): Web-accessible gene expression data analysis. *Bioinformatics* **18**, 1540-1541 (2002a).

Colantuoni, C., Henry, G., Zeger, S., and Pevsner, J. Local mean normalization of microarray element signal intensities across an array surface: Quality control and correction of spatially systematic artifacts. *Biotechniques* **32**, 1316-1320 (2002b).

Dopazo, J., Zanders, E., Dragoni, I., Amphlett, G., and Falciani, F. Methods and approaches in the analysis of gene expression data. *J. Immunol. Methods* **250**, 93-112 (2001).

Eickhoff, B., Korn, B., Schick, M., Poustka, A., and van der Bosch, J. Normalization of array hybridization experiments in differential gene expression analysis. *Nucleic Acids Res.* **27**, e33 (1999).

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863-14868 (1998).

Golub, T. R., et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537 (1999).

Haverty, P. M., et al. HugeIndex: A database with visualization tools for high-density oligonucleotide array data from normal human tissues. *Nucleic Acids Res.* **30**, 214-217 (2002).

Hegde, P., et al. A concise guide to cDNA microarray analysis. *Biotechniques* **29**, 548-550, 552-554, 556 passim (2000).

Hilsenbeck, S. G., et al. Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *J. Natl. Cancer. Inst.* **91**, 453-459 (1999).

Iyer, V. R., et al. The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83-87 (1999).

Kaufman, L., and Rousseeuw, P. J. *FINDING GROUPS IN DATA.*

An Introduction to Cluster Analysis, Wiley, New York, 1990.

Khan J., et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58**, 5009–5013 (1998).

Kuruvilla, F. G., Park, P. J., and Schreiber, S. L. Vector algebra in the analysis of genome-wide expression data. *Genome Biol.* **3**, 11.1–11.11 (2002).

Landgrebe, J., Wurst, W., and Welzl, G. Permutation-validated principal components analysis of microarray data. *Genome Biol.* **3**, 19.1–19.11 (2002).

Liao, B., Hale, W., Epstein, C. B., Butow, R. A., and Garner, H. R. MAD: A suite of tools for microarray data management and processing. *Bioinformatics* **16**, 946–947 (2000).

Manduchi, E., et al. Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics* **16**, 685–698 (2000).

Misra, J., et al. Interactive exploration of microarray gene expression patterns in a reduced dimensional space. *Genome Res.* **12**, 1112–1120 (2002).

Motulsky, A. G. Jewish diseases and origins. *Nat. Genet.* **9**, 99–101 (1995).

Olshen, A. B., and Jain, A. N. Deriving quantitative conclusions from microarray expression data. *Bioinformatics* **18**, 961–970 (2002).

Perou, C. M., et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA* **96**, 9212–9217 (1999).

Perou, C. M., et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).

Quackenbush, J. Computational analysis of microarray data. *Nat. Rev. Genet.* **2**, 418–427 (2001).

Raychaudhuri, S., Stuart, J. M., and Altman, R. B. Principal components analysis to summarize microarray experiments: Application to sporulation time series. *Pac. Symp. Biocomput.* 455–466 (2000).

Raychaudhuri, S., Sutphin, P. D., Chang, J. T., and Altman, R. B. Basic microarray analysis: Grouping and feature reduction. *Trends Biotechnol.* **19**, 189–193 (2001).

Sawitzki, G. Quality control and early diagnostics for cDNA microarrays. *R News* **2**, 6–10 (2002).

Schuchhardt, J., et al. Normalization strategies for cDNA microarrays. *Nucleic Acids Res.* **28**, E47 (2000).

Sherlock, G. Analysis of large-scale gene expression data. *Brief. Bioinform.* **2**, 350–362 (2001).

Shipp, M. A., et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **8**, 68–74 (2002).

Singer, C. The Fasciculo di Medicina Venice 1493. R. Lier and Co., Florence, 1925.

Smid-Koopman, E., et al. Gene expression profiles of human endometrial cancer samples using a cDNA-expression array technique: Assessment of an analysis method. *Br. J. Cancer* **83**, 246–251 (2000).

Spellman, P. T., et al. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* **3**, 46.1–46.9 (2002).

Tamayo, P., et al. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96**, 2907–2912 (1999).

Tusher, V. G., Tibshirani, R., and Chu, G. Significance analysis of microarray applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. Systematic determination of genetic network architecture. *Nat. Genet.* **22**, 281–285 (1999).

Wall, M. E., Dyck, P. A., and Brettin, T. S. SVDMAN–singular value decomposition analysis of microarray data. *Bioinformatics* **17**, 566–568 (2001).

West, M., et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. USA* **98**, 11462–11467 (2001).

Wittes, J., and Friedman, H. P. Searching for evidence of altered gene expression: A comment on statistical analysis of microarray data. *J. Natl. Cancer. Inst.* **91**, 400–401 (1999).

Zolman, J. F. *Biostatistics*. Oxford University Press, New York, 1993.

70 CHEMISTRY OF THE PROTEIDS CHAP.

	1	2	3	4	5	6	7
	Globin of the Oxyhæmoglobin of Horse's Blood.	Serum-albumin of Horse's Blood.	Serum-globulin.	Egg-white.	Egg-albumin crystallised.	Albumin of Yolk.	Caseinogen (see p. 307 for other products).
Glycocoll	0[9]	0[10]	3·52[96]	...	...	...	0[43 96]
Alanin	4·19[9]	2·68[10]	2·22[96]	...	...	...	0·9[96]
Leucin	29·04[9]	20·48[10]	18·7[96]	22·6[40]	...	...	10·5[96]
Phenylalanin	4·24[9]	3·08[10]	3·84[96]	+[3]	...	...	3·2[96]
a-Pyrrolidin-carboxylic acid	2·34[9]	1·04[10]	2·76[96]	+[3]	...	...	3·2[1 96]
Glutaminic acid	1·73[9]	1·52[10]	2·20[96]	+[40]	+[48]	...	10·7[96]
Aspartic acid	4·43[9]	3·12[10]	2·54[96]	+[37]	+[48]	...	1·2[96]
Cystin	0·31[9]	2·53[43]	1·51[43]	0·4[43]	0·29[43]	...	0·065
Serin	0·56[9]	0·6[10]	...	...	...	...	0·48[14]
Oxy-a-Pyrrolidin-carboxylic acid	1·04[9]	...	...	...	...	...	0·25[14]
Tyrosin	1·33[9]	2·1[10]	...	0·58[65]	1·5[31]	...	4·5[96]
Lysin	4·28[9]	...	...	+[18 75]	...	...	5·8[96]
Histidin	10·96[9]	...	...	+[18]	+[18]	+[18]	2·6[96]
Arginin	5·42[9]	...	...	+[17]	+[16 17]	+[17]	4·84[96]
Tryptophane	+[9]	+[10]	...	+[80]	...	...	1·5[84]
Ammonia	0·93[50]	1·2[49]	1·75[49]	...	1·5[49]	...	1·8[96]
Cystein	...	+[44]	...	+[44]	...	...	0[44]
Amino-valerianic acid	...	...	...	...	...	...	1·[96]
Glucosamin	...	...	...	...	10–11[34]	...	0[67]
Diamino-trioxydodecanoic acid	...	...	...	...	...	...	0·75[96]

尽管对于我们而言，蛋白质是由20种氨基酸组成的已是一个不争的事实，然而生活在19世纪末的化学家们还是在努力地想弄清楚蛋白质的组成成分。在那时，只有为数不多的几种蛋白质为人们所知道，这些蛋白质包括所谓的清蛋白（包括血清蛋白、乳球蛋白、纤维蛋白原、肌浆球蛋白、组蛋白）、狭义的蛋白质（proteids、血红蛋白、黏液素）和硬蛋白（如胶原质、角蛋白、硬性蛋白和淀粉质）。在这些蛋白质中只有极少的几种我们能够得到它们的晶体［如从马身上得到的血红蛋白和血清蛋白、卵清蛋白、鱼卵磷蛋白（大马哈鱼中的清蛋白）］。古斯塔夫·曼恩（Gustav Mann，1906，第70～第75页）第一次描述了蛋白质的基本组分，他利用一些方法将51种蛋白质的混合物解离成不可再用相同的方法解离的单位。上表显示了其中7种蛋白质（见上表的列）的解离结果，表中的行代表在解离蛋白质时发现的各种组分。在这些组分中，大部分都是氨基酸，如glycocoll（胶糖）是glycine（甘氨酸）以前的名字。这张表还说明很早的时候蛋白质就可以被分析了，科学家们付出了巨大的努力，希望不仅了解单个蛋白质的性质还希望了解来自不同物种中相关蛋白质的联系

第 8 章

蛋白质分析和蛋白质组学

8.1 引言

生命体主要由 5 种物质组成，它们分别是蛋白质、核酸、脂类、水和碳水化合物。在这 5 种物质中，蛋白质是决定细胞性质最主要的物质。人们在建造房屋时需要一张蓝图，然后根据蓝图确定怎样建造房屋，DNA 经常被比喻为构筑生命体的蓝图，相应地，蛋白质就是构筑生命体最主要的材料。蛋白质在生命过程中发挥着巨大的作用，它们执行着大部分生物功能。这些功能包括结构功能（如细胞骨架中的肌动蛋白）、酶功能（很多蛋白质可以催化生物反应，常见的蛋白质催化功能是使生物反应加速一定数量级），以及在细胞内或细胞间转运物质的功能。如果把 DNA 类比成建造房屋时所用的蓝图，那么蛋白质就是建成后房屋的主体部分，它不仅是房屋的墙和地板，还包括房屋的管道系统、电力产生系统、传输系统以及垃圾回收系统。

> 蛋白质是多肽聚合物，它是由一系列按顺序排列的氨基酸残基构成的。"protein" 这个词是瑞典化学家 Jöns Jacob Berzelius（1779～1848）在 19 世纪 30 年代引入的。"protein" 在希腊语中的意思是"最重要的"。

8.2 如何认识蛋白质：从四个视角看蛋白质

这一章的开始部分将从四个不同的视角认识蛋白质（总结见图 8-1）。这四个视角分别为：

① 蛋白质家族；

② 蛋白质的物理性质；

③ 蛋白质的细胞定位；

④ 蛋白质的生物功能。

认识蛋白质的第一个视角是蛋白质家族，在这一部分中，我们会定义诸如家族（family）、结构域（domain）、模体（motif）等术语。然后再讨论蛋白质的物理性质和评估蛋白质物理性质的方法。

第三个和第四个视角，即蛋白质的细胞定位和生物功能，对应于 Gene Ontology 协会定义的一组概念，我们将从 Gene Ontology 的角度认识蛋白质这两方面的性质。在过去的若干年里，成千上万的 DNA 序列被测定，其中包括多种生物的全基因组序列。大量序列被测定

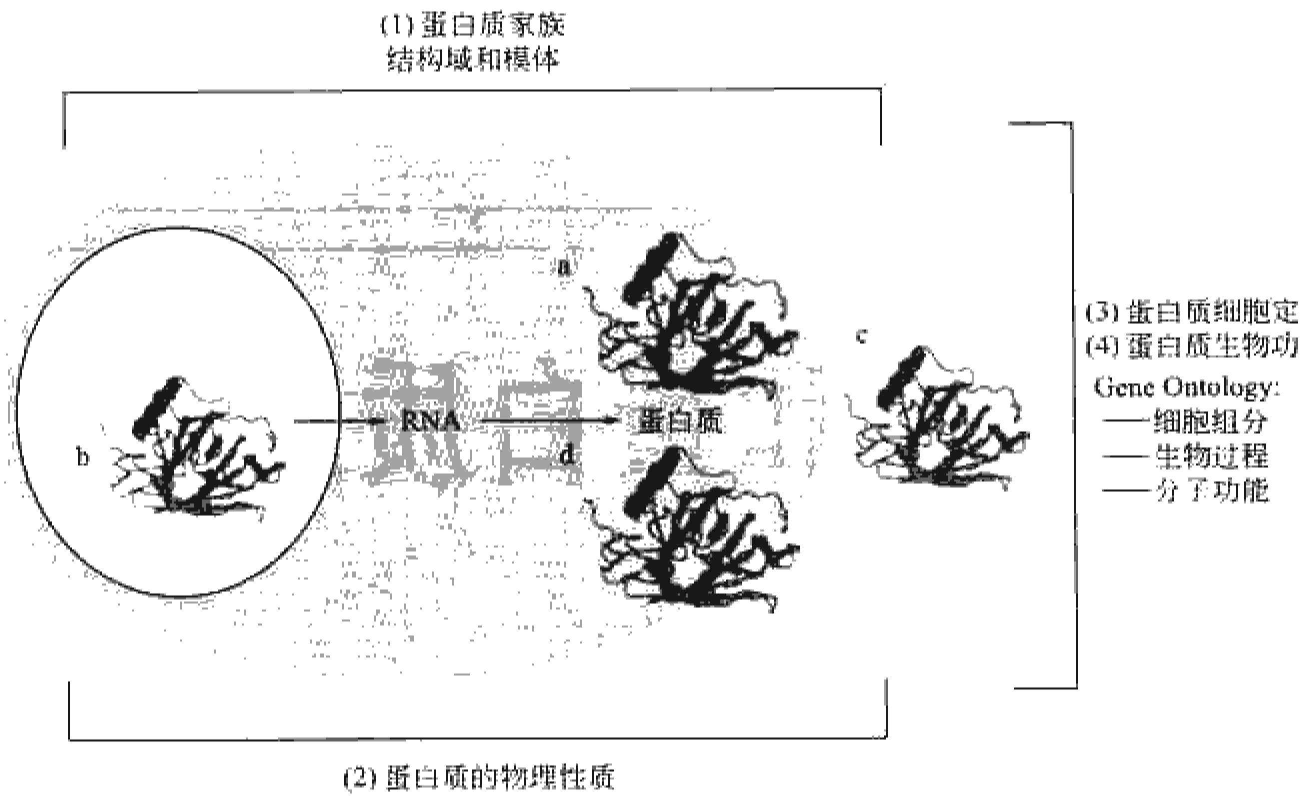


图 8-1 对蛋白质的概括。一个蛋白质是由一系列氨基酸残基组成的，这一组氨基酸残基的顺序是由一个特定的基因决定的。蛋白质可以根据不同方面的性质分为很多类，包括家族、定位、物理性质和功能。(1) 蛋白质家族是根据蛋白质的同源性来定义的；同一个家族内的蛋白质之间，可能仅有一部分序列是同源的。一些蛋白质家族和模体数据库（在第10章讨论）将成百上千的蛋白质根据同源性进行了分类，每一个分类中的蛋白质功能上具有相关性。(2) 可以对蛋白质的物理性质进行描述，例如大小（分子量）、形状（如半径和摩擦系数）、电荷（等电点），以及对翻译后修饰（见正文部分）或是因为 mRNA 的选择性剪切和蛋白质水解作用造成的蛋白质异构产物的描述。(3) Gene Ontology (GO) 协会根据细胞组分、生物过程（如转录或内吞）和分子功能（例如催化或转运）将蛋白质进行分类，一个蛋白质可能属于多个分类类别。GO 系统提供了一组动态而又可控的词汇，它们可以用于所有的真核蛋白。如这张图中所显示的，一个蛋白质可能出现在细胞的不同位置上：它有可能是一个可溶性蛋白，存在于胞质中（标签 a），或者存在于一个细胞器里，如细胞核中（标签 b），或是被分泌到胞外（标签 c）。一个蛋白质还有可能被固定在细胞表面的质膜上（标签 d）。蛋白质可能通过两种方式定位在质膜上：跨膜区结构域和表面接触

带给了生物信息学家一个挑战，那就是如何从这些序列中找到基因，然后给基因加上注释，即给这些基因提供关于它们性质或功能的简单描述。为了能为基因和基因的产物进行一致而又合理的注释，Gene Ontology 协会确立了一个基因注释的框架，在这个框架中定义了一组灵活而又可控的词汇，这些词汇从如下 3 个方面描述蛋白质的性质和功能：它们在细胞中充当什么组分（cellular component），它们在哪个生化过程中起作用（biological process），以及它们起怎样的分子功能（molecular function）。在介绍这一部分的时候，我们先对 Gene Ontology 工程进行一个整体的介绍，然后再介绍与蛋白质细胞定位和生物功能对应的 Gene Ontology 分类，即细胞组分（cellular component）和分子功能（molecular function）。

通过上面四个视角，可以窥测单个蛋白质的性质。除此之外，在这一章中还要探索一些其他方法，这些方法并不针对单个蛋白，而是针对一个基因组中的所有蛋白（proteome，即蛋白质组）或特定通路（pathway）中的蛋白质相互作用。第 9 章将介绍多条同源序列比对的方法，并给出一些收集了已比对的蛋白序列的数据库的网址，如 Pfam 和 InterPro。我们还要将多序列比对的结果以系统发生树的形式显示出来。

8.2.1 视角 1. 结构域和模体：蛋白质的模块性质

让我们从几种类型的蛋白质开始对结构域的讨论。最简单的一类蛋白质是，这种蛋白质在目前所有已知的数据库中均没有发现同源序列。这种情况可能经常出现，但是一个基因组编码的一半或一半以上的蛋白质都没有同源序列的情况是不常见的。即使一个蛋白质没有已知的同源序列，它的其他性质（如跨膜区结构域、磷酸化位点、预测出的二级结构等）也会给我们了解该蛋白质的结构或功能提供一些线索。

另一类蛋白质是确实有直系同源序列或旁系同源序列的蛋白质。这种蛋白质至少能找到一条同源序列，且两条序列存在具有显著相似性的区域。这些有显著序列相似性或显著结构特征的区域有很多名称，如签名（signature）、结构域（domain）、模块（module）、模块元件（modular element）、折叠子（fold）、模体（motif）、模式（pattern）或重复（repeat）。

这些术语的定义虽然有所不同，但都和多条进化上相关的蛋白质序列所共有的区域有关。为了统一起见，本书中主要采用 InterPro 数据库提供的术语定义。InterPro 是一个整合了多个蛋白质家族、结构域和功能位点数据库的文档资源。

> InterPro 数据库和它的成员数据库（PRINTS、PROSITE、Pfam、ProDom、SMART 和 TIGRFAMS）将在这一章和第 10 章中讨论。可以通过 http://www.ebi.ac.uk/interpro/ 来访问 InterPro 数据库。

签名（signatures）的概念很宽广，它确定一个蛋白质分类，可能指结构域（domain）、家族（family）或模体（motif）。仅考虑单独的一个蛋白质时，我们仅能得到很少一部分关于其结构和功能的信息；但是将它与相关序列比对找到保守部分后，从保守序列中可以推测出很多信息。签名主要可以分为两类，每一类签名都可以用各自的方法确定。

① 结构域是蛋白质中能折叠成特定三维结构的一段区域（Doolittler，1995）。结构域也能被称为模块（Henikoff 等，1997；Sonnhammer 和 Kahn，1994）。在叙及蛋白质的三维结构时，还经常用折叠子（fold）这个术语。一组拥有相同结构域的蛋白被称为一个蛋白质家族。结构域根据细胞定位（如出现在胞质中的结构域被称为胞内结构域，定位在细胞外的结构域被称为胞外结构域）或空间结构（如结合二价锌离子的锌指结构域）又可再做区分。

目前，已有很多有关蛋白质家族的数据库，如 Pfam 和 SMART。有关蛋白质家族数据库的内容我们还将在第 10 章中详细介绍。表 8-1 和表 8-2 列出了 InterPro 和 SMART 数据库对家族（family）、结构域（domain）、重复（repeat）以及相关术语的定义。

表 8-1 InterPro 数据库对蛋白质家族及相关术语的定义

术 语	定 义
家族(family)	InterPro 数据库定义一组进化上相关的共享一个或多个结构域/重复区的蛋白为一个家族。InterPro 数据库中"type=family"的登录条目也包括了那些代表一个家族的短的保守区域，并不要求保守区域要覆盖整个蛋白质
结构域(domain)	InterPro 数据库中的结构域是指一个独立的结构单元，它们可能单独存在也可能与其他结构域或重复区相连。结构域也是进化上相关的序列。InterPro 数据库中"type=domain"可以确定一段序列是否属于一种结构域，并没有精确的定义结构域的边界
重复区(repeat)	InterPro 数据库中定义的重复区并不单独折叠成一个球状的结构域。例如，要折叠成一个球状结构域需要 6～8 个拷贝的 *WD40* 重复区。"type=repeat"登录条目中还包括一些短的重复模体序列，它们并不折叠成一个球状结构域
翻译后修饰(posttranslational modification)	翻译后修饰包括诸如 N-糖基化位点等序列模体。这些序列模体的定义是细胞中可以被分子识别的序列区域。按照这些序列区域组织起来的蛋白质分组并不代表蛋白质之间存在进化上的联系

注：摘自 http://www.ebi.ac.uk/interpro/user.manual.html。

表 8-2 SMART 数据库中对蛋白质结构域和模体的定义

术 语	定 义
结构域(domain)	保守的结构单元,包含独特的二级结构组合和疏水内核。在小的富含二硫键和 Zn^{2+} 或 Ca^{2+} 结合的结构域中,疏水内核往往分别由形成二硫键的两个半胱氨酸和金属离子提供。具有相同功能的同源结构域往往具有序列上的相似性
结构域成分(domain composition)	具有相同结构域成分的蛋白质是指至少含有一个拷贝的我们提交的查询结构域的蛋白质
结构域组织形式(domain organization)	含有相同结构域组织形式的蛋白质是指含有我们提交查询的结构域(也可以包含其他结构域),且这些结构域在蛋白质中的排列顺序与我们提交的顺序一致
模体(motif)	序列模体是指短的保守的多肽段。含有相同模体的蛋白质并不一定是同源的
轮廓(profile)	一个轮廓就是一张打分的表。表中包含对每个氨基酸残基位置的打分,还有对空格的罚分。它表示一个蛋白质家族,可以用来搜索序列数据库

注:1. SMART 是一个可以对用户提交的蛋白质序列中可能含有的结构域进行自动识别和注释的工具(见第10章)。
2. 摘自 http://smart.embl-heidelberg.de/help/smart.glossary.shtml。

② 模体(motif,或称指纹,fingerprint)是蛋白质序列中较短的保守区域(下文还有讨论)。模体通常是指按一定模式排列的氨基酸残基,它通常决定一个蛋白质家族(Bork 和 Gibson,1996)。模体的长度一般是10~20个氨基酸残基,尽管实际中的模体有可能更长或更短。一些简单而常见的模体在一组蛋白质中发现并不意味着这组蛋白质是同源的,例如形成跨膜区结构域或保守磷酸化位点的模体。而另一些情况中,小的模体则是一个蛋白质家族的标志。

为了更加直观,我们介绍一些具体的结构域的例子,表8-3列出了人类基因组中编码的蛋白质中最常见的一些结构域。第14~第16章也给出了类似的列表,其中有编码在其他物种基因组中的更为丰富的蛋白质结构域的信息。一般说来,如果两个蛋白质拥有一个相同的结构域,那么这两个蛋白具有相同或相似的功能。例如,已发现的数百个小GTP结合蛋白都拥有一个相同的结构域,其中一些成员被认为是通过与GTP的结合与解离调节胞内运输膜泡的对接和融合(Geppert 等,1997);另一些分子量相对较小的成员则在细胞周期调控和细胞骨架形成时起作用。这个蛋白质超家族被组织成相关的亚家族,同一个亚家族中的蛋白质被认为拥有相同的功能。

表 8-3 人类中15个最常见的结构域

InterPro ID	基因组中的拷贝数	蛋白质数目	名 字
IPR000822	30034	1093	锌指,C2H2 类型
IPR003006	2631	1032	免疫球蛋白/主要组织相容性复合物
IPR000561	4985	471	EGF 类似结构域
IPR001841	1356	458	锌指,RING
IPR001356	2542	417	同源异型框
IPR001849	1236	405	Pleckstrin 类似结构域
IPR000504	2046	400	RNA 结合区 RNP-1(RNA 识别模体)
IPR001452	2562	394	SH3 结构域
IPR002048	2518	392	钙结合 EF-hand
IPR003961	2199	300	纤粘连蛋白,类型Ⅲ
IPR001478	1398	280	PDZ/DHR/GLGF 结构域
IPR005225	261	261	小 GTP 结合蛋白结构域
IPR000210	583	236	BTB/POZ 结构域
IPR001092	713	226	碱性螺旋-环-螺旋(bHLH)二聚结构域
IPR002126	5168	226	钙黏着蛋白

注:来源于欧洲生物信息学研究所(EBI)蛋白质组分析站点(http://www.ebi.ac.uk/proteome)(2002年8月),基于 InterPro 数据库(http://www.ebi.ac.uk/interpro/index.html)。

现在，让我们把注意力集中到单个结构域上来，蛋白质之间共享一个结构域的方式可以有多种（图 8-2）。整个蛋白质可以仅仅由一个结构域组成，例如 lipocalin 结构域 [图 8-2(a)]。还有很多其他小的球蛋白也是由单个结构域构成的。

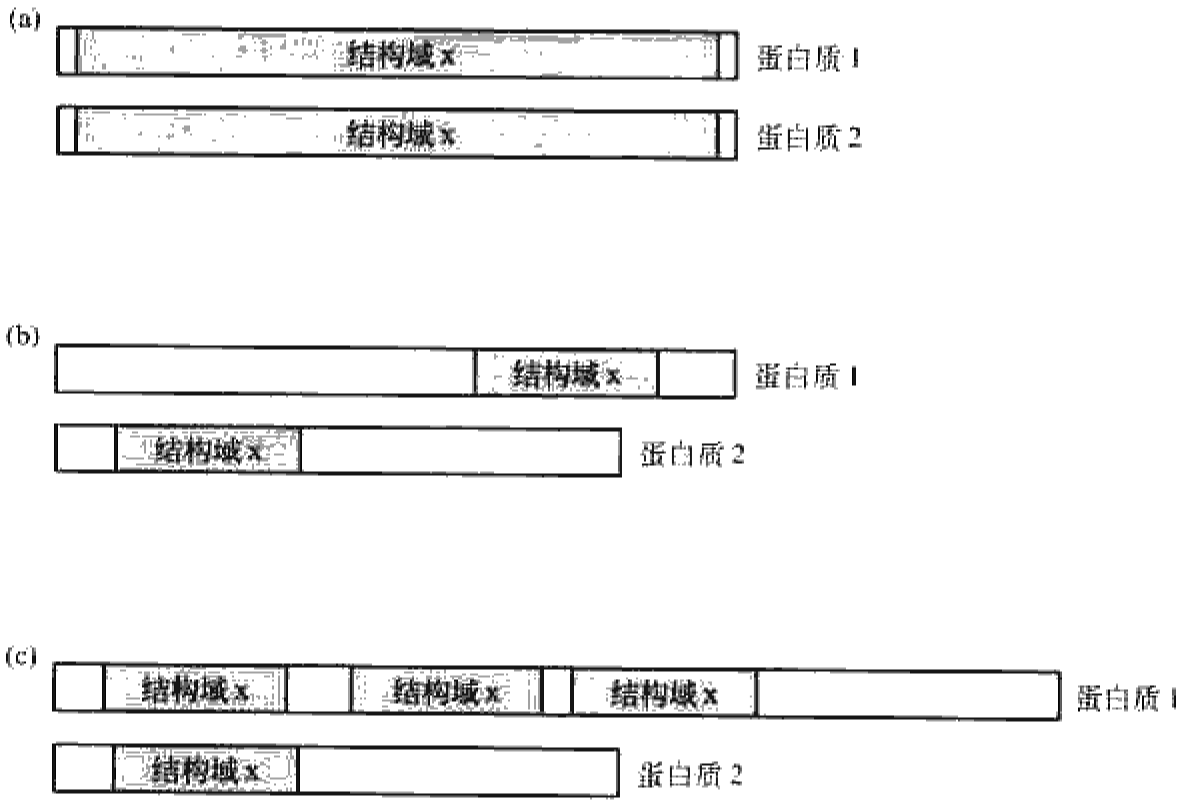


图 8-2 蛋白质可以通过多种方式共享一个结构域。(a) 一个结构域可能占据了一条蛋白质的整条序列，如 lipocalin 家族的蛋白质。(b) 结构域可能是一组高度相关的氨基酸序列，但它们只充当整条蛋白质序列一个亚基。与甲基化的 DNA 结合的转录因子家族就是一个这样的例子（见图 8-3）。(c) 结构域也可能在一个蛋白质中重复很多次（有时有很多拷贝数）。这样的结构域在同源蛋白质中可能出现任意次。一个例子就是含有 fibronectin Ⅲ-like 重复的蛋白质家族（见图 8-4）

另一种更为常见的情况是，蛋白质由多个结构域构成，每个结构域只充当蛋白质的一个亚基。比较两个蛋白质序列时经常发现，相同的结构域在不同的蛋白质中位于不同的区域 [图 8-2(b)]。一个例子就是有 6 个蛋白质拥有同一个结构域，这个结构域使这 6 个蛋白都具有结合甲基化的 DNA 的能力，但这 6 个蛋白质中的同一个结构域分别位于序列的不同区域上。甲基-CpG-结合蛋白 2（Mecp2）就是这 6 个蛋白质中的一个。它是一个转录抑制因子，可以结合在很多基因的上游调控区（MECP2 基因的突变会导致 Rett 综合征或女孩的神经逻辑错乱，它还是女性智力障碍的几个常见原因之一，见框 18-2）。我们可以用 Blastp 软件搜索与 Mecp2 蛋白有显著相似性的序列，这可以揭示序列中哪一段是保守的结构域。图 8-3(a) 所示的是 BLAST 软件的格式化参数输入页面中的一部分，从中可以看出甲基-CpG-结合结构域（MBD）在多个蛋白质结构域数据库中都存在。BLAST 的搜索结果显示 Mecp2 蛋白只有一部分与其他 5 个 MBD 蛋白 [图 8-3(b)] 序列匹配。进一步分析由这 5 个蛋白组成的 MeCP2/MBD 家族，会发现这 5 个蛋白大小相差很大，只有 MBD 结构域相互匹配，其他部分并不匹配。

蛋白质家族是怎么定义的呢？是否一组仅共有一个结构域的同源蛋白质就可以被称为一个蛋白质家族呢？上面的例子中，5 个蛋白中的 MBD 结构域显然是同源的（来自于同一祖先）；虽然它们除了 MBD 结构域外没有显著相似的区域，但是这组蛋白仍旧构成一个蛋白质家族。蛋白质家族的定义就是一组进化上相关的蛋白质，这些蛋白质共享一个或多个结构域。

第三种情况是一个蛋白质仅含一种结构域，但这种结构域在蛋白中可能重复出现多次 [图 8-2(c)]。两个最常见的例子就是人类中的免疫球蛋白结构域（immunogloblin domain）和纤连蛋白结构域（fibronection domain）。这两个结构域经常出现在有胞外区的蛋白中。有一点值得

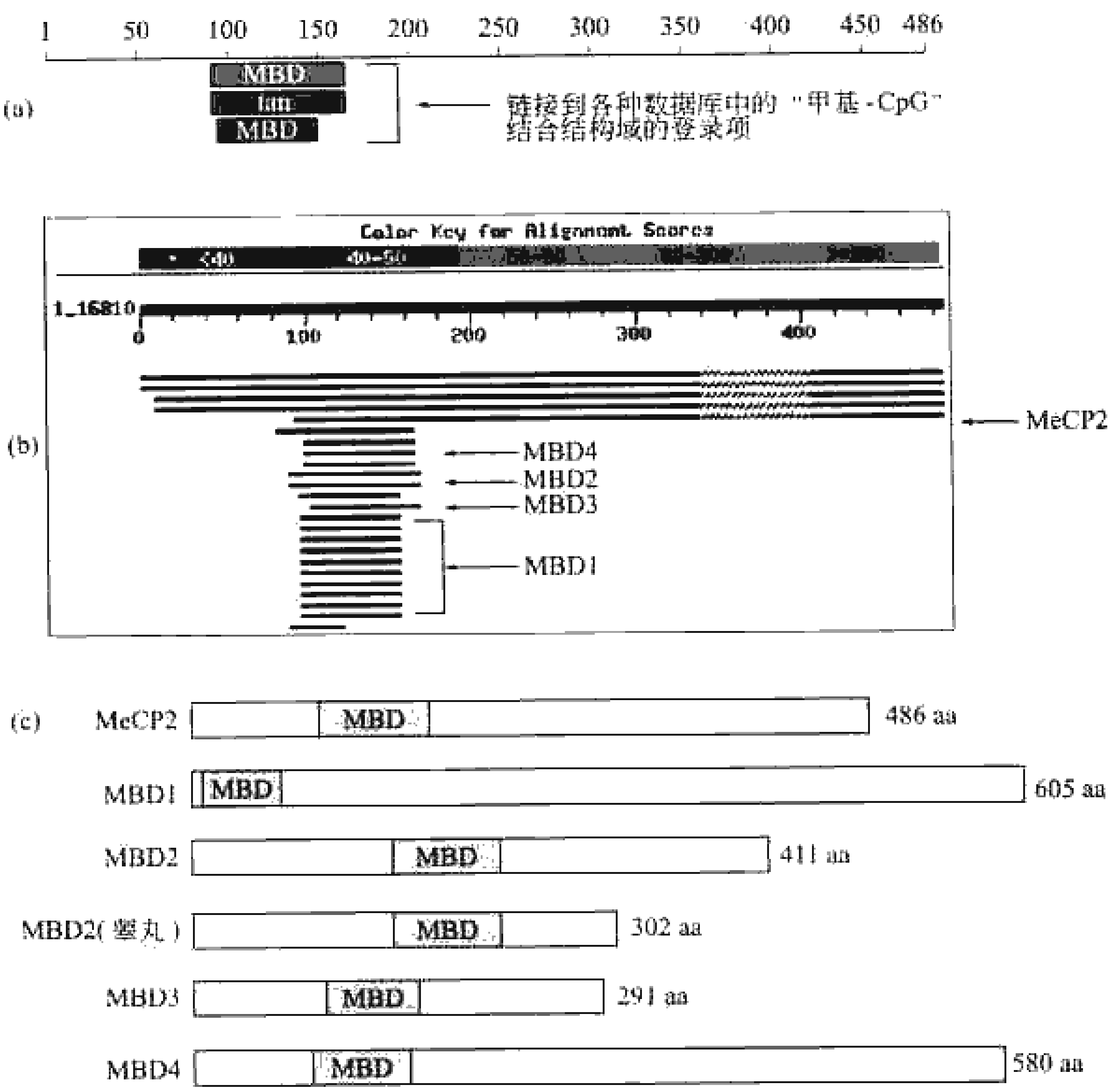


图 8-3 甲基结合结构域（methy-binding domain）在很多人类蛋白中都存在。为了显示这个结构域，我们向 Blastp 软件提交了甲基-CpG-结合蛋白 2（methyl-CpG-binding protein 2，简称 MeCP2）的序列，将搜索的范围限制为人类蛋白质。(a) BLAST 软件的格式化页面显示 MeCP2 这个蛋白包含的结构域也被其他几个数据库收录了。(b) BLAST 搜索结果页面显示有 5 个独立提交的 MeCP2 序列与提交的序列匹配（最上面的 5 条比对结果）。此外，MeCP2 中还有一个含有大约 80 个氨基酸残基的区域与其他类型的甲基-CpG-结合蛋白匹配，它们分别是 MBD1（NP.056671）、MBD2（NP.003918）、一个睾丸特异的 MBD2 的异构体（NP.056647）、MBD3（NP.003917）和 MBD4（NP.003916）。(c) 这些蛋白质的大小差别很大，并且结合甲基化 DNA 的结构域也出现在蛋白质的不同位置上。从 BLAST 的匹配结果看，这些蛋白质除了甲基化 DNA 结构域以外没有其他序列上有显著相似性的区域了

注意，即这种或类似的其他胞外结构域在人类或多细胞线虫（如秀丽线虫）中大量存在，但在单细胞真核生物酵母中却几乎不存在（Copley 等，1999）。比较各种生物基因组中编码的蛋白质家族的差异可以为我们了解不同生物的生命过程提供一些有用的线索。

8.2.2 多结构域蛋白的复杂性

到目前为止我们仅考虑了蛋白质仅含一种结构域的情况，而一个蛋白质含有多种结构域的情况更复杂也更常见。HIV-1 的 pol 蛋白就是一个这样的例子（Frankel 和 Young，1998）（见图 8-7）。*pol* 基因编码了一条很长的多肽链，这个多肽链将裂解成 3 个独立的具有不同生物学活性的蛋白质，这 3 个蛋白质分别是天冬氨酸蛋白酶（aspartyl protease）、反

转录酶（以 RNA 为模板的 DNA 聚合酶）和整合酶（integrase）。注意 HIV-1 的 pol 蛋白是一个多结构域蛋白裂解成多个单结构域蛋白的例子。其他的多结构域蛋白（如图 8-4 所示的免疫球蛋白）在成熟的多肽链中仍保持着相互独立且又相连的结构域。

为了进一步研究 pol 蛋白的序列，我们需要借助蛋白质专家分析系统（Expert Protein Analysis System，简称 ExPASy，见图 8-5）。首先进入序列提取系统（SRS），然后搜索

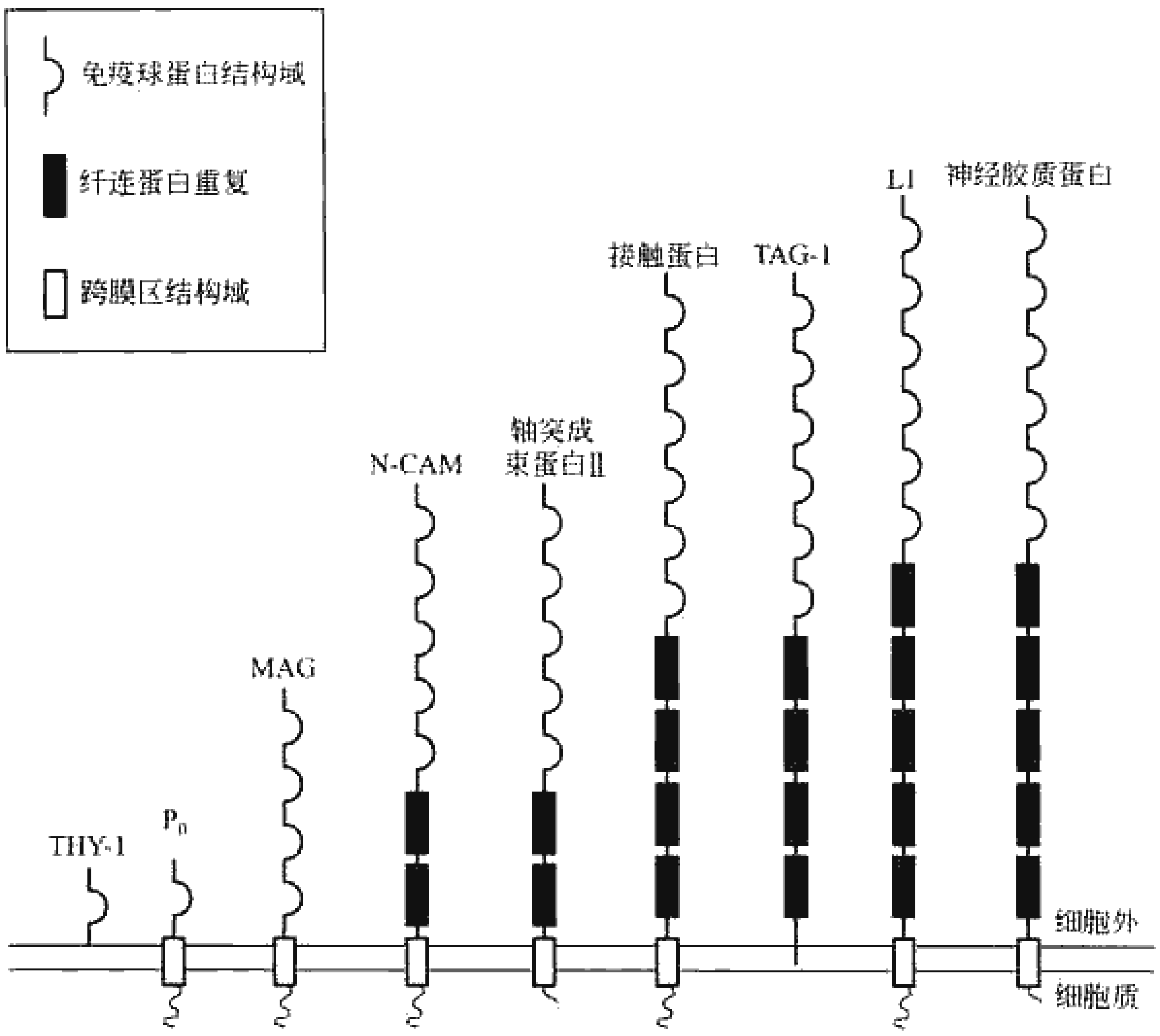


图 8-4 很多结构域在蛋白质中有多个拷贝，两个最常见的例子就是免疫球蛋白结构域（immunoglobulin domains）和纤连蛋白重复区（fibronectin repeat）。这些结构域在蛋白质的胞外区极其常见

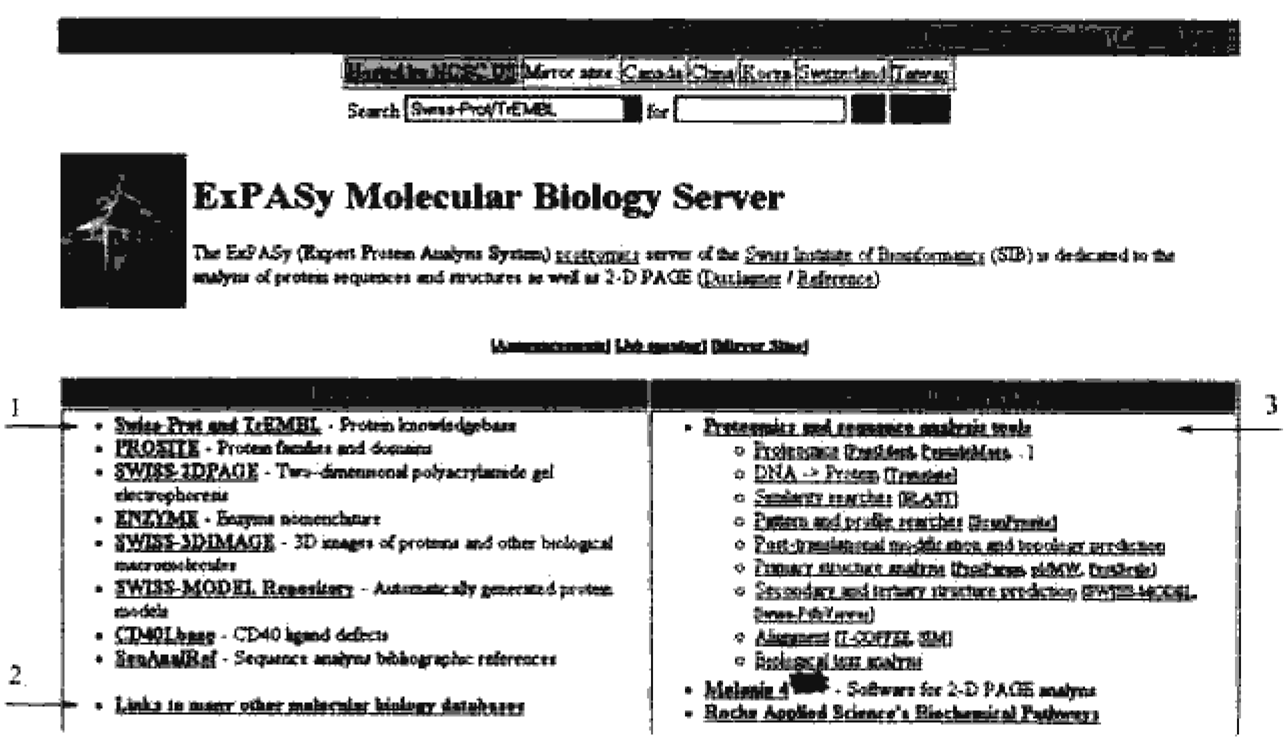


图 8-5 ExPASy 是最主要的一个基于网页的蛋白质分析服务器（http://www.expasy.ch/）。这个站点提供了一个序列提取系统的接口（箭头 1）和一个很大的并且组织得很好的数据库链接表（箭头 2）。此外这个站点还提供了很多蛋白质分析工具（箭头 3）

HIV-1（物种）和pol。选择有1003个氨基酸残基的蛋白序列（SwissProt数据库的accession序号是P03369）。SwissProt数据库的这条记录中包含很多相关数据库的链接，其中6个是链接到蛋白质家族数据库的（或结构域或模体数据库），它们分别是InterPro、Pfam、PROSITE、ProDom、BLOCKS和DOMO（见下文）。ProDom后面的链接连向一个页面，这个页面中有一系列至少与HIV-1的pol蛋白共享一个相同结构域的蛋白质。点击该链接将显示如图8-7所示的结果，图中显示多条蛋白质与HIV-1的pol蛋白共有至少一个结构域。

ExPASy SRS的网址是 http://www.expasy.ch/srs5/。我们在本章中介绍的很多工具都可以在ExPASy网站上获得。ExPASy是由瑞士生物信息学研究所维护的。

PROSITE数据库的网址是 http://www.expasy.org/prosite/。PROSITE数据库中定义的"轮廓"（profile）是指基于统一的语法对蛋白质模体的定量描述。"模式"是指用类似正则表达式的语法（如下文的例子）对模体进行的定性描述。模体是指符合一个特定模式或轮廓的生物序列。关于这些定义可以参见 ftp://www.expasy.org/databases/prosite/release/profile.txt。

8.2.3 蛋白质模式：模体或蛋白质的指纹特征

在一个结构域中会有一些短而连续的保守氨基酸残基。这些短的保守的氨基酸片段被称为模体（或指纹）。一个模体的例子就是酶活性位点附近的氨基酸序列。在HIV-1 pol蛋白的天冬氨酸蛋白酶（aspartyl proteases）结构域中，天冬氨酸残基（asp）对酶的催化活性至关重要。PROSITE数据库收集了几乎所有目前已知的蛋白质模体，它可以充当查阅蛋白质模体的字典（Sigrist等，2002）。点击SwissProt数据库页面后面交叉索引部分的链接（图8-6）或直接访问PROSITE站点都可以找到有关天冬氨酸蛋白酶（aspartl proteases）模体的信息（图8-8）。天冬氨酸蛋白酶模体由12个氨基酸残基构成：[LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]-x-[LIVMFSTNC]-x-[LIVMFGTA]。这种格式和PHI-BLAST（见第5章）所用的格式相同。

模体是蛋白质结构域的一部分。在几乎所有的lipocalins中都可以找到一个短的模体GxW。PROSITE数据库定义的lipocalins的保守氨基酸模式（PROSITE document是PDOC00187）在GXW氨基酸残基周围新增了一些其他残基，它定义的模体是[DENQ]-x-[DENQGSTARK]-x(0,2)-[DENQARK]-[LIVFY]-{CP}-G-{C}-W-[FYWLRH]-x-[LIVMTA]。GxW序列在上面的模体中是用G-{C}-W表示的，大括号表示除了C以外其他的氨基酸残基都可能出现在这个位置上。很多模体非常短也非常常见，例如很多激酶底物丝氨酸（serine）或苏氨酸（threonime）残基附近的序列。这样的模体并不一定是一个蛋白质家族所特有的，它们可能在很多非同源的蛋白质中出现。在PROSITE站点搜索关键词"kinase"，会出现多个登录条目，它们既包括激酶中的签名（signatures）也包括激酶底物中的签名。其中的一个登录条目链接的页面显示的是蛋白激酶C（PKC）保守磷酸化位点的信息：[ST]-x-[RK]（S或T是磷酸化位点，x表示任意氨基酸残基）（PROSITE document的编号是PDOC00005）。这个简单的模体在蛋白质中出现过的次数达到千余次。

很重要的一点是，PROSITE数据库定义的正则表达式（或者模式）是定性（如是匹配

Cross-references	
EMBL	K02087, AAB59876.1, - [EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	A03968, GNVWA2
PDB	3HVP; 15-JAN-90 [ExPASy / RCSB] 1CPI; 08-MAR-96 [ExPASy / RCSB] 1MTR; 01-AUG-96 [ExPASy / RCSB] 1YTG; 12-MAR-97 [ExPASy / RCSB] 1YTH; 12-MAR-97 [ExPASy / RCSB]
HIV	K02087, POL$SF2
MEROPS	A02.001; -
InterPro	IPR001995; Asp_prot_retrov. IPR001969; Asp_protease. IPR001037; Integrase_C. IPR003308; Integrase_zn. IPR002156; RNaseH. IPR000477; RVTse. IPR001584; Rve. Graphical view of domain structure.
Pfam	PF00552; integrase; 1. PF02022; Integrase_Zn; 1. PF00075; rnaseH; 1. PF00665; rve; 1. PF00077; rvp; 1. PF00078; rvt; 1.
PROSITE	PS00141; ASP_PROTEASE; 1. PS50175; ASP_PROT_RETROV; 1.
ProDom	[Domain structure / List of seq. sharing at least 1 domain].
BLOCKS	P03369.
DOMO	P03369.
ProtoMap	P03369.
PRESAGE	P03369.
DIP	P03369.
SWISS-2DPAGE	GET REGION ON 2D PAGE.

图 8-6　SwissProt 数据库 HIV-1 pol（accession P03369）的登录页面含有很多链接到其他蛋白质结构域数据库的链接。本章中我们将要对这些蛋白质结构域数据库进行探讨，包括 InterPro、Pfam 和 PROSITE。我们可以在 ExPASy 站点的序列提取系统（http://www.expasy.ch/srs5/）上搜索获取类似的 SwissProt 数据库的登录页面。箭头表示有 6 个不同的数据库含有 HIV-1 pol 中的结构域和模体信息

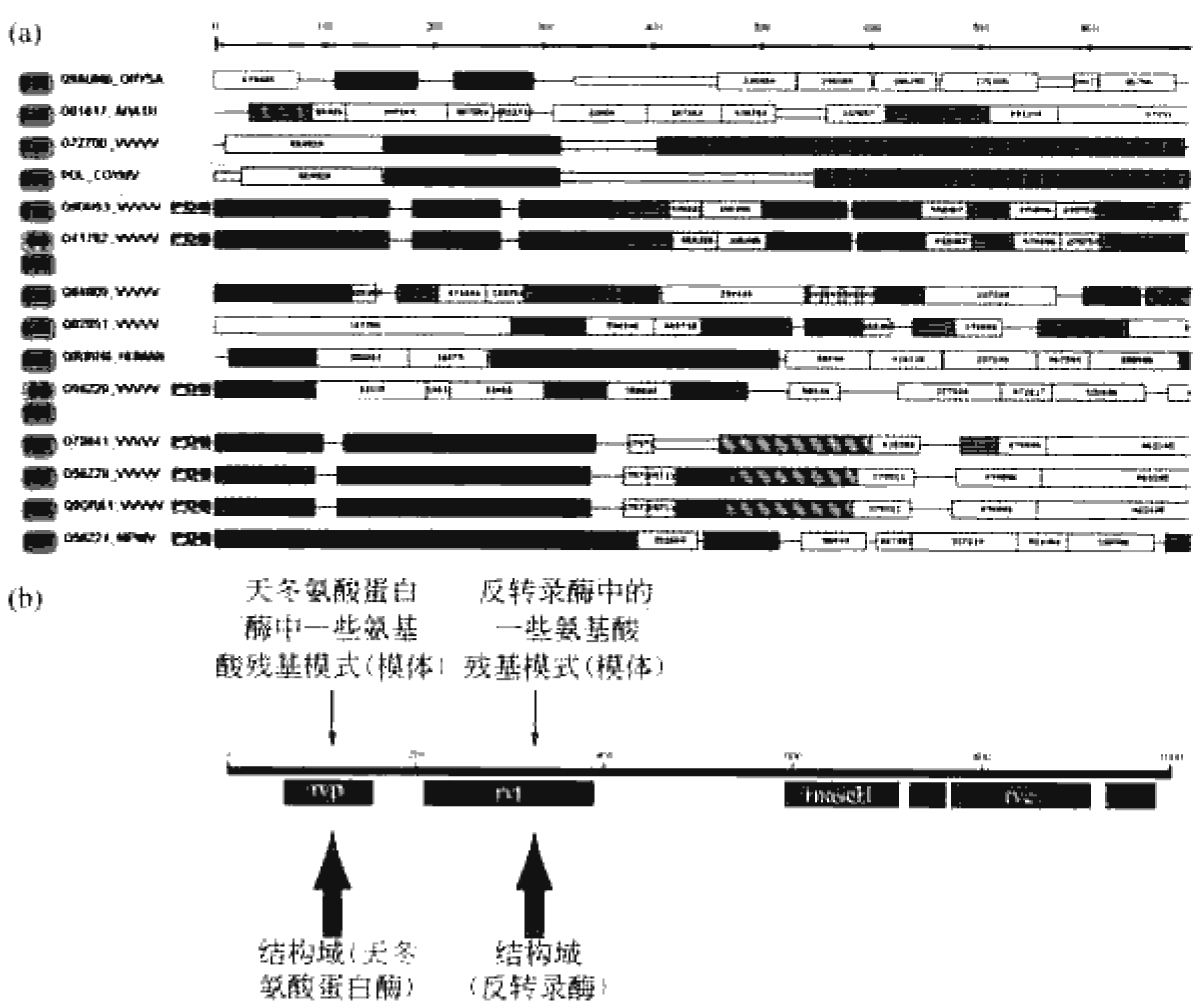


图 8-7　(a) 点击 SwissProt 数据库 HIV-1 pol 登录页面上的 ProDom 链接会出现一个页面，这个页面中显示了数百个蛋白质与 pol 共有一个结构域。也可以直接访问 ProDom 数据库（http://prodes.toulouse.inra.fr/prodom/doc/prodom.html）来获取这个页面。结构域在蛋白质序列中充当不同的模块，图中只列出了 14 个蛋白质序列。(b) 蛋白质可能含有相对较大的结构域和通常仅由几个氨基酸残基构成的模式（模体）。尽管模式或模体不构成一个已知的三维构象，但它包含的氨基酸残基却可能是一个蛋白质家族的特征序列

PROSITE cross-reference(s)

PS00141, ASP_PROTEASE
PS50175, ASP_PROT_RETROV

Documentation

Asparty1 proteases, also known as acid proteases, (EC 3.4.23.-) are a widely distributed family of proteolytic enzymes [1,2,3] known to exist in vertebrates, fungi, plants, retroviruses and some plant viruses. Aspartate proteases of eukaryotes are monomeric enzymes which consist of two domains. Each domain contains an active site centered on a catalytic aspartyl residue. The two domains most probably evolved from the duplication of an ancestral gene encoding a primordial domain. Currently known eukaryotic aspartyl proteases are:

- Vertebrate gastric pepsins A and C (also known as gastricsin).
- Vertebrate chymosin (rennin), involved in digestion and used for making cheese.
- Vertebrate lysosomal cathepsins D (EC 3.4.23.5) and E (EC 3.4.23.34).
- Mammalian renin (EC 3.4.23.15) whose function is to generate angiotensin I from angiotensinogen in the plasma.
- Fungal proteases such as aspergillopepsin A (EC 3.4.23.18), candidapepsin (EC 3.4.23.24), mucoropepsin (EC 3.4.23.23) (mucor rennin), endothiapepsin (EC 3.4.23.22), polyporopepsin (EC 3.4.23.29), and rhizopuspepsin (EC 3.4.23.21).
- Yeast saccharopepsin (EC 3.4.23.25) (proteinase A) (gene PEP4). PEP4 is implicated in posttranslational regulation of vacuolar hydrolases.
- Yeast barrierpepsin (EC 3.4.23.35) (gene BAR1); a protease that cleaves alpha-factor and thus acts as an antagonist of the mating pheromone.
- Fission yeast sxa1 which is involved in degrading or processing the mating pheromones.

Most retroviruses and some plant viruses, such as badnaviruses, encode for an aspartyl protease which is an homodimer of a chain of about 95 to 125 amino acids. In most retroviruses, the protease is encoded as a segment of a polyprotein which is cleaved during the maturation process of the virus. It is generally part of the pol polyprotein and, more rarely, of the gag polyprotein.

Conservation of the sequence around the two aspartates of eukaryotic aspartyl proteases and around the single active site of the viral proteases allows us to develop a single signature pattern for both groups of protease. A profile was developed to specifically detect viral aspartyl proteases, which are missed by the pattern.

Description of pattern(s) and/or profile(s)

Consensus pattern	[LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]-x-[LIVMFSTNC]-x-[LIVMFGTA] [D is the active site residue]
Sequences known to belong to this class detected by the pattern	ALL
Other sequence(s) detected in Swiss-Prot	37
Sequences known to belong to this class detected by the profile	ALL viral- type proteases
Other sequence(s) detected in Swiss-Prot	3
Note	these proteins belong to families A1 and A2 in the classification of peptidases [4,E1]

图 8-8 PROSITE 数据库 aspartyl proteases（包含了 HIV-1 pol）的登录页面。这个数据库描述了超过 1000 个蛋白质家族的签名（模体）

还是不匹配）而非定量的。当然模式（patterns）还可以有更复杂的定义，如在一个给定位置上可能出现的氨基酸残基，与给定的模式相比哪些不匹配是不被允许的等。其他许多数据库，例如 Pfam、ProDom、SMART 和 TIGRFAMs（将在第 10 章介绍）使用了轮廓（profiles）来代替这种严格定义的模式（patterns）。轮廓和模式有一点相似，就是都是以多序列比对为基础的；但轮廓采用一个分值矩阵来表示不同位置的特异性，这也让轮廓描述的蛋白质序列的长度增加了。

要了解哪些站点提供了模体分析工具，可以参见“8.7 网络资源”中的表 8-11。

8.2.4　视角 2. 蛋白质的物理性质

蛋白质有很多固有的物理性质，这些物理性质可能是由蛋白本身的氨基酸序列造成的，也可能是大量的翻译后的修饰造成的（表 8-4）。有些修饰可以让疏水基团与蛋白质共价结合以使蛋白质能插入脂质双分子层。这样的例子有棕榈酰化（palmitoylation）、法尼基化（farnesylation）、豆蔻酰基化（myristoylation）和肌醇糖脂修饰（inositol glycolipid attachment）（图 8-9）。

表 8-4　蛋白质的一些物理性质

性　质	经典方法	例　子
氨基酸模体	—	PDZ 结构域（如一氧化氮合成酶），螺旋卷曲结构域（如血凝素、突触融合蛋白、SNAP-25、肌球蛋白）
等电点	通过等电聚焦实验得到	—
分子量	通过分子半径和沉降系数得到	—
翻译后修饰：磷酸化	酶分析	突触蛋白
翻译后修饰：糖基化	酶分析	神经生长因子，神经细胞附着分子
翻译后修饰：异戊二烯化	—	Lamin B，G 蛋白 γ 亚基，rab3A
翻译后修饰：棕榈酰化		β 肾上腺素受体，GAP-43，胰岛素受体，rhodopsin，nAChR
翻译后修饰：豆蔻酰化		PKA，Gα 亚基，MARCKS 蛋白，钙调神经磷酸酶
翻译后修饰：GPI 锚定蛋白	酶分析	碱性磷酸酶，thy-1，朊蛋白，5′核苷酸酶，尿调理素
沉降系数	通过蔗糖梯度离心实验得到	—
分子半径	通过凝胶电泳实验得到	—
跨膜区结构域	通过亚细胞成分分离实验得到	—

注：G 蛋白为鸟苷三磷酸（GTP）结合蛋白；GAP-43 为分子质量为 43ku 的生长相关蛋白（growth-associated protein of 43 ku）；MARCKS 为丝氨酸化的富含丙氨酸的 C 激酶底物（myristoylated alanine-rich C-kinase substrate）；nAChR 为 N-乙酰胆碱受体（nicotinic acetylcholine receptor）；PDZ 结构域，其中 P 代表突触后膜致密蛋白 PSD-95，D 代表果蝇肿瘤抑制因子 discs-large，Z 代表紧密连接蛋白 ZO-1；PKA 为蛋白激酶 A；SNAP-25 为分子质量为 25ku 的突触体维系蛋白（synaptosomal-associated protein of 25ku）；Rab3A 为大鼠大脑 GTP 结合蛋白 3A；thy-1 为胸腺细胞 1。

目前拥有大量的基于网页（web）的服务可用于预测和评价蛋白质的物理性质。我们只需要输入一条蛋白质序列便可通过这些服务预测出蛋白质的一些物理性质，如分子量（mass）、等电点（p*I*；图 8-10 显示的是预测的结果，表 8-14 中列出了可用的 web 资源）、氨基酸的组成成分、糖基化位点（表 8-15）、磷酸化位点（表 8-16）和酪氨酸硫化位置（图 8-11）等。

如何评价各种预测程序预测出来的结果呢？它们是否准确呢？目前，主要还只能通过在实验室做实验的方式来评价蛋白质的物理性质，这种方式每次只能处理一种蛋白质（Cooper，1977）。蛋白质的分子量可以通过凝胶过滤色谱或聚丙烯酰胺凝胶电泳（PAGE）的方

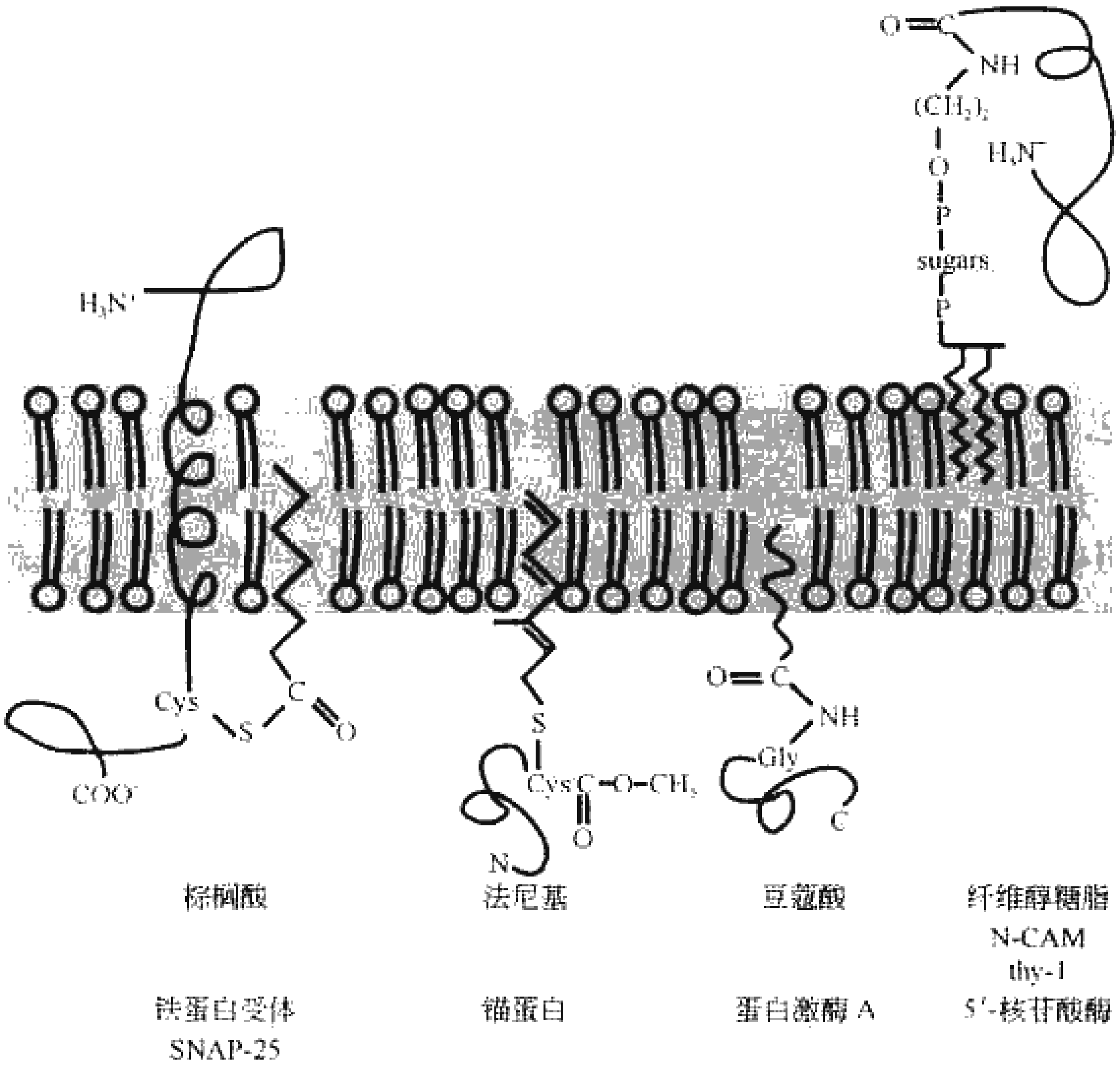


图 8-9 蛋白质翻译后会被进行很多的修饰，例子有棕榈酰化（palmitoylation）（如加到运铁蛋白受体和 SNAP-25）、法尼基化（farnesylation）（如加到锚蛋白）、豆蔻酰基化（myristoylation）（如加到蛋白激酶 A）和肌醇糖脂修饰（inositol glycolipid attachment）（如神经细胞支持分子，thy-1 和 5′核苷酸酶）。尽管这些共价修饰可以通过生化方法进行研究，但仍旧有很多网站提供了蛋白质共价酸修饰位点的预测

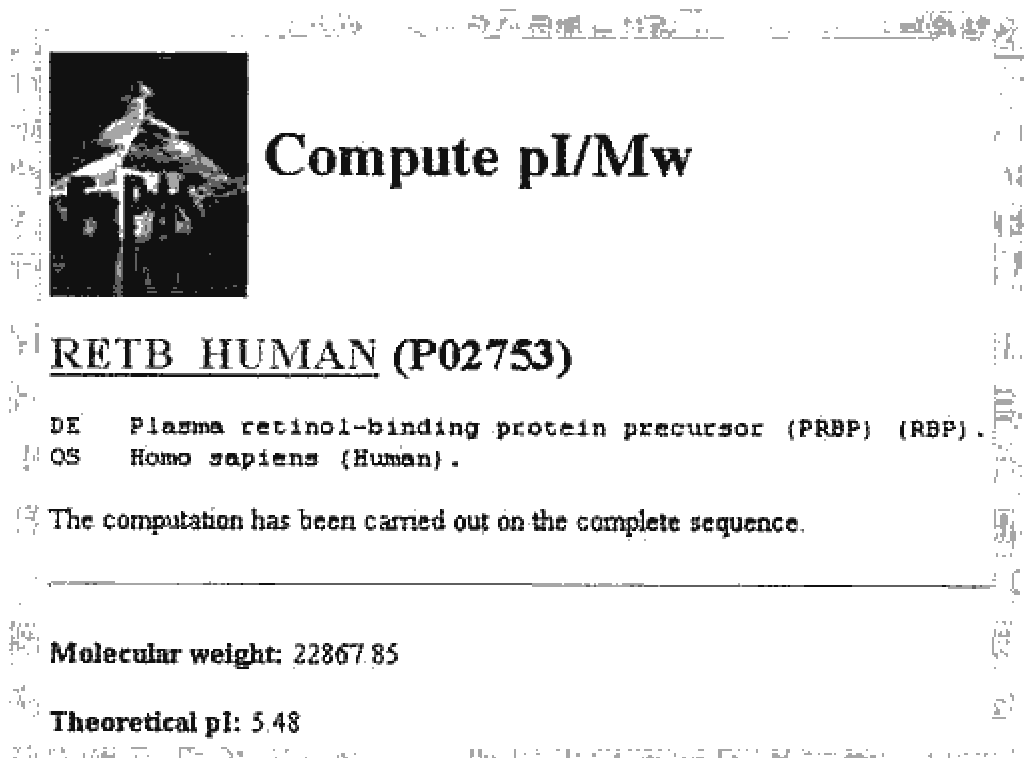


图 8-10 ExPASy 网站提供的计算 pI/M_r 的服务，它可以计算提交序列的等电点和分子量，图中显示的是计算出来的甲基醇结合蛋白的等电点和分子量

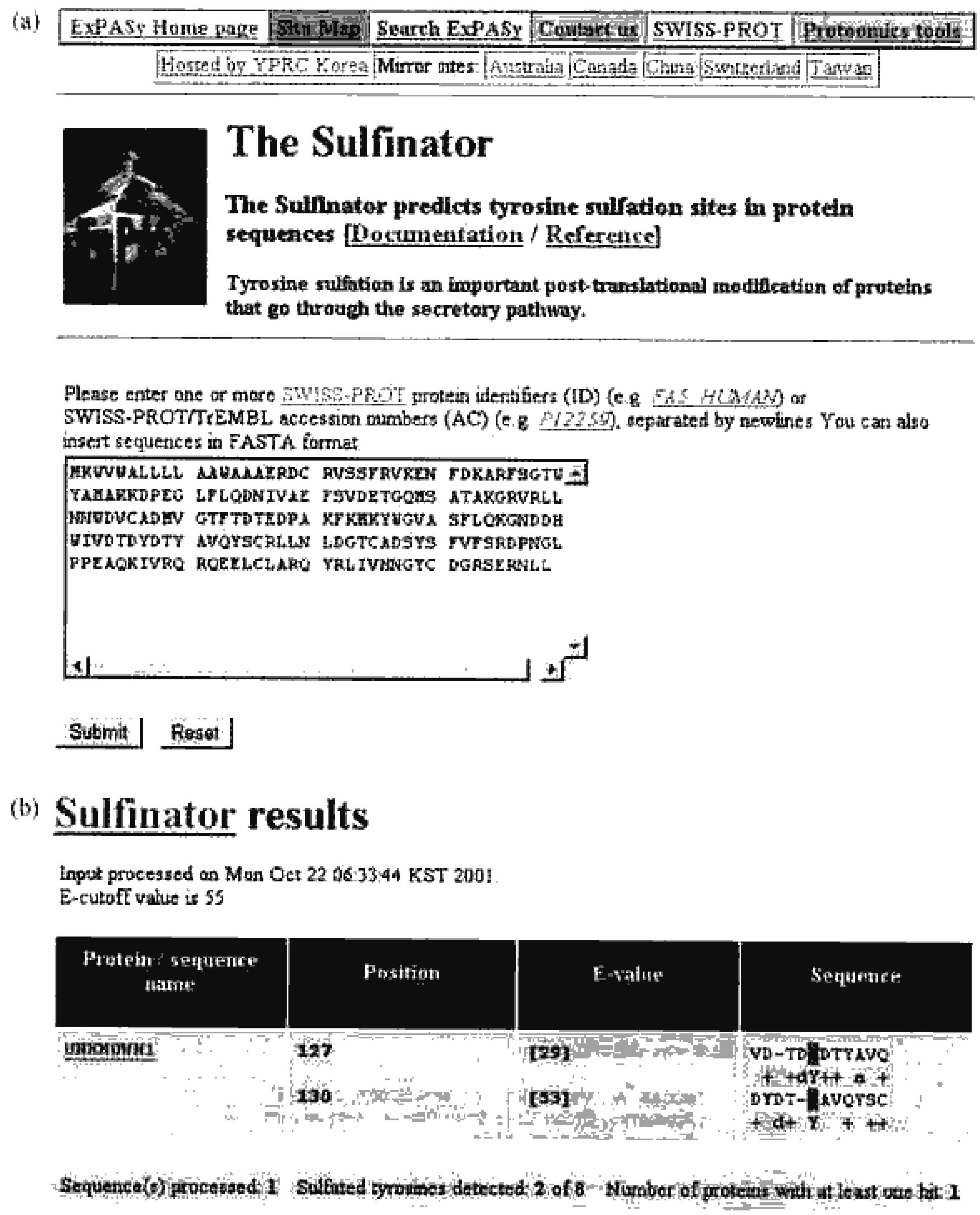


图 8-11 ExPASy 站点提供了很多蛋白质分析工具，如 Sulfinator。(a) 蛋白质序列是输入信息。(b) 输出结果中有酪氨酸残基上可能会出现的硫化位点。这些关于硫化、磷酸化、糖基化或是其他翻译后修饰位点的预测是我们设计实验检测蛋白质功能的基础

法来估计。它们的形状可以通过摩擦系数推测出来，而摩擦系数则可以通过凝胶过滤和蔗糖梯度离心两个实验的结合来计算得到。然而上述技术不能同时运用于大量的蛋白质。现在，几乎所有的用生物信息学工具研究的蛋白质都还没有通过实验的筛选，它们仅仅是从基因组 DNA 序列或 cDNA 序列中被预测出来。

基于蛋白质一级序列的预测程序的可信度要高于那些基于蛋白质其他信息的预测程序，如一些二级结构性质，包括拓扑结构和翻译后修饰。对于那些具有典型氨基酸组成的蛋白质，对它们的分子量和等电点（p*I*，见图 8-10）的预测结果是很准确的。这些性质也可以通过实验的方法来验证，如凝胶电泳和等电聚焦。预测程序可以很准确地确定蛋白质中一段短的特定的序列是否是磷酸化或硫化位点的保守序列，但是在活体细胞中这些位点不一定就真的会被修饰。

预测蛋白质跨膜区拓扑结构的程序的准确性如何呢？我们可以很容易利用搜索工具搜索到一个预测的结果。但是，从根本上讲，预测跨膜区拓扑结构是一个细胞生物学的问题，它需要

用到细胞生物学的工具去找到明确的答案。很多蛋白质都含有10～25个连续的疏水氨基酸残基，这些氨基酸残基很可能构成的跨膜结构域。检测疏水氨基酸残基是否构成跨膜结构域的最严格的方法是通过实验来验证，如免疫细胞化学（immunocytochemistry）。从兔子、小鼠或其他动物体内提取出含有某种特异抗体的血清，然后用这些血清可以检测固定在显微镜载玻片上的样品中的抗原（如细胞表面的受体）。对于细胞膜不可渗透的细胞，抗血清可以让细胞膜上的蛋白质朝向胞外的部分显示出来。但如果细胞可以被去垢剂渗透，则抗血清不仅显示胞外部分的蛋白质，胞内（胞质）部分的蛋白质也会被显示出来。诸如此类的细胞生物学的分析手段经常被用来确定一个蛋白质有几个跨膜结构域，而且很多情况下，实验的结果和通过氨基酸的疏水性绘图（hydropathy plots）所得预测结果（如Ratnam等，1986）相矛盾。

类似地，很多程序可以预测糖基化位点、磷酸化位点或其他特殊的位点。这些预测结果能给生物学家研究蛋白质的翻译后修饰提供很有意义的帮助，但是这些结果还需要进一步的实验验证。

在第9章中还将详细讨论蛋白质的结构。很多程序可以预测蛋白质的二级结构（见第9章表9-1）。其中有一种二级结构被称为螺旋卷曲结构（coiled-coil regions），这种结构与蛋白质-蛋白质相互作用结构域有密切的关系（Lupas等，1991；Lupas，1997）（图8-12和图8-13）。

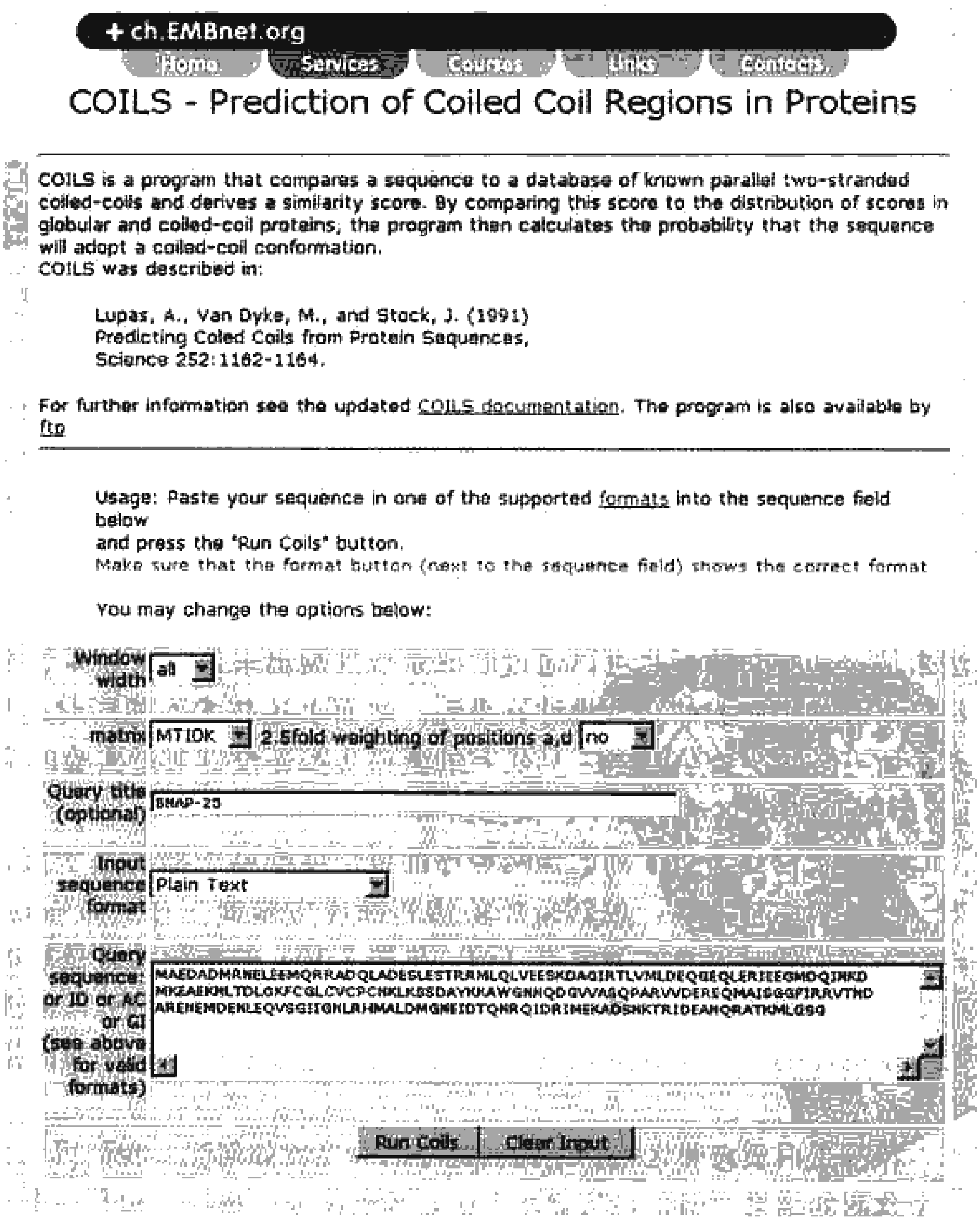


图8-12 coils程序可以确定一条蛋白质序列是否形成螺旋卷曲结构

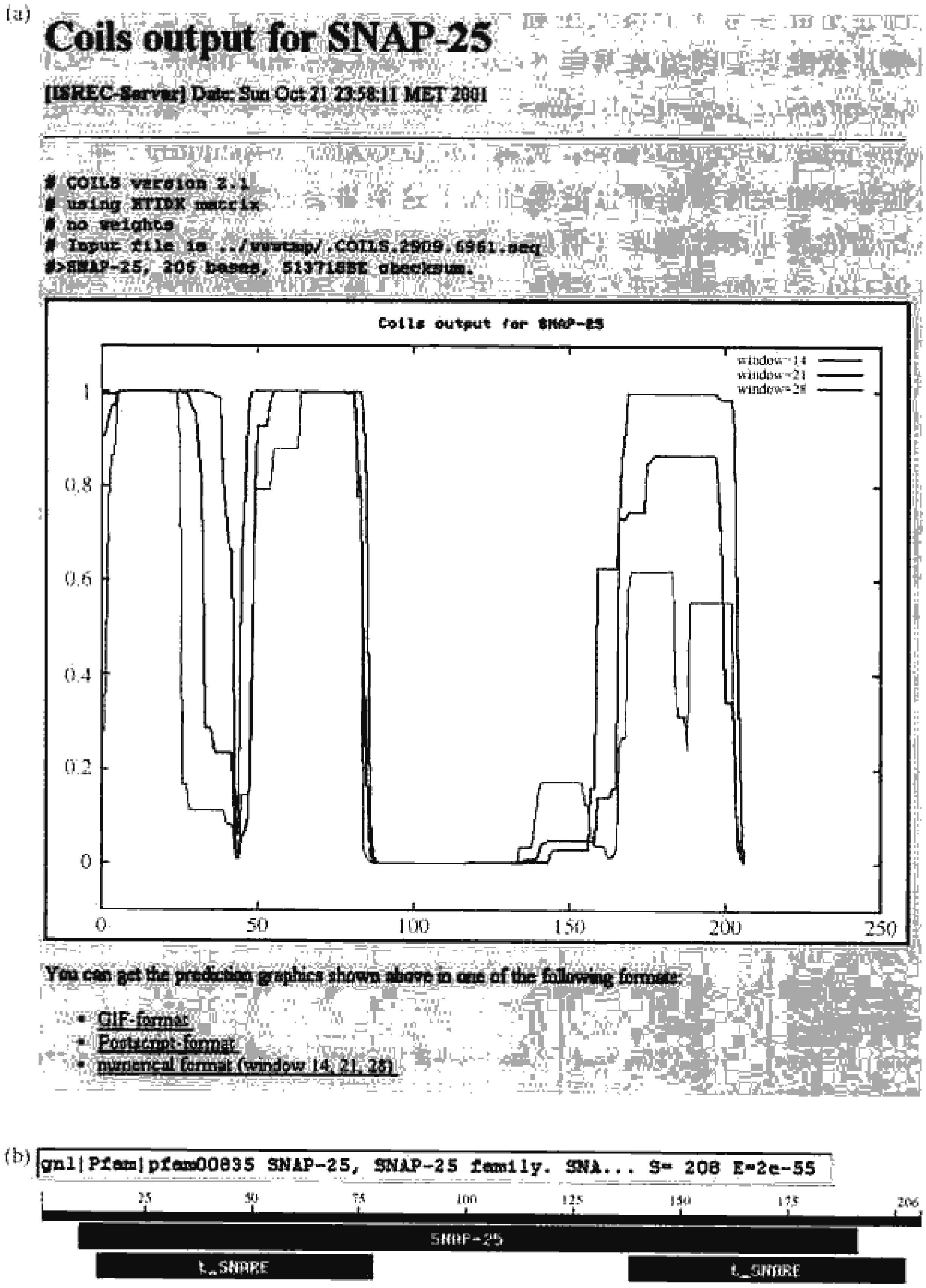


图 8-13 (a) ISREC 服务器上 coils 程序的输出结果。输入的信息是 SNAP-25 (NP. 003072) 序列。结果显示的是蛋白质在整个序列长度内 (x 轴) 形成螺旋卷曲二级结构的概率 (y 轴)。使用了 3 种窗体大小来计算概率。螺旋卷曲结构经常出现在蛋白质相互作用结构域中。在这个例子中，SNAP-25 是一个与质膜外表面结合的蛋白质，它的螺旋卷曲结构使它能够紧密地与其他两个蛋白结合来调整神经突触前膜上突触囊泡的对接和神经递质的释放。(b) 将 SNAP-25 的序列提交给 blastp 程序，可以找到 SMART 数据库中螺旋卷曲结构域 (coiled-coil 结构域)。coils 网站地址为 http://www.isrec.isb-sib.ch/cgi-bin/print_hit_bold.pl/software/COILS_form.html

8.2.5　视角 3 和视角 4 的介绍：Gene Ontology 协会

Ontology 是一个描述性的概念。Gene Ontology (GO) 协会致力于这样一项工程：编辑一组动态的而又可控的词汇来描述基因和基因产物（主要是蛋白质）不同方面的

性质。这个协会是由参与3个模式生物数据库建设的科学家发起的，这三个模式生物数据库分别是芽殖酵母基因组数据库（*saccaromyces* genome database，SGD）、果蝇基因组数据库（*drosophila* genome database，简称 FlyBase）和小鼠基因组信息数据库（mouse genome information database，简称 MGD/GXD）（Ashburner 等，2000，2001）。紧接着，其他物种相关的数据库也陆续加入了这个协会（表 8-5）。GO 数据库不是以其自身为中心而是依靠外部数据库（例如小鼠数据库），这些外部数据库中收录的基因及其产物都将用 GO 定义的词汇进行注释。因此 GO 是与时俱进与相互合作的代表，它致力于统一基因及其产物注释的方式。目前我们可以登录几个站点（表 8-6）搜索和查询 GO 中定义的词汇（或者称术语）。此外，NCBI（见第2章）位置链接（LocusLink）中也包含了 GO 词汇。

表 8-5 加入了 Gene Ontology 协会的数据库和组织

数据库或组织	物种学名	物种常用名	数据库 URL
Berkeley Drosophila Genome Project（伯克利果蝇基因组计划）	果蝇（*Drosophila melanogaster*）	fly（果蝇）	http://www.fruitfly.org/
Compugen DictyBase	黏菌（*Dictyostelium discoideum*）	slime mold（黏菌）	http://www.cgen.com/
European Bioinformatics Institute（欧洲生物信息学研究所，简称 EBI）	多种生物		http://ebi.ac.uk/
FlyBase	果蝇（*Drosophila melanogaster*）	fly（果蝇）	http://flybase.bio.stanfor.edu
Genome Knowledge Base（GKB）at Cold Spring Harbor Laboratory（冷泉港实验室的基因组知识数据库）	多种生物	—	http://genomeknowledge.org/
Mouse Genome Database（MGD）and Gene Expression Database（GXD）（小鼠基因组数据库和基因表达数据库）	小鼠（*Mus musculus*）	mouse（小鼠）	http://www.informatixs.jax.org
Pathogen Group at the Wellcome Trust Sanger Institute	多种生物	—	http://www.sanger.ac.uk/Teams/Pathogen/
PomBase	裂殖酵母（*Schizosaccharomyces cerevisiae*）	fission yeast（裂殖酵母）	http://www.sanger.ac.uk/Projects/S.pombe
Rat Genome Database（RGD）（大鼠基因组数据库）	大鼠（*Rattus*）	rat（大鼠）	http://rgd.mcw.edu/
Saccharomyces Genome Database（SGD）（芽殖酵母基因组数据库）	芽殖酵母（*Saccharomyces cerevisiae*）	baker's yeast（芽殖酵母）	http://genome-www.stanford.edu/Saccharomyces/
The Arabidopsis Information Resource（TAIR）	阿拉伯芥（*Arabidopsis thaliana*）	thale cress（阿拉伯芥）	http://www.arabidopsig.org/
The Institute for Genomic Research（TIGR）（基因组研究所）	多种生物	—	http://www.tigr.org/
WormBase	蠕虫（*Caenorhabditis elegans*）	worm（蠕虫）	http://www.wormbase.org/

注：摘自 http://www.geneontology.org/。

表 8-6　可以获得 Gene Ontology 数据的有用站点

显示 GO 的程序	描　述	站点 URL
AmiGO	GO 浏览器源自伯克利果蝇基因组计划	http://www.godatabase.org/cgi-bin/go.cgi
Mouse Genome Informatics (MGI) GO Browser(小鼠基因组 GO 浏览器)	由 Jackson 实验室开发维护	http://www.informatics.jax.org/userdocs/GO.help.shtml
EBI 的"Quick GO"	由欧洲分子生物学实验室(EMBL)和欧洲生物信息研究中心(EBI)开发，整合了 InterPro 数据库(见第 10 章)	http://www.ebi.ac.uk/ego/manual.html
Expression Profiler(EP) GO Browser(表达谱 GO 浏览器)	GO 浏览器和分析工具是欧洲生物信息研究所 EP 项目的一部分	http://ep.ebi.ac.uk/EP/
Cancer Gene Anatomy Project(CGAP)(癌症相关基因浏览器)GO Browser	由美国国立卫生研究院的国家癌症研究中心开发维护	http://cgap.nci.nih.gov/Genes/AllAboutGO/

GO 有 3 个主要的方面，分别是分子功能、生物过程和细胞组分。分子功能主要指单个基因产物分子所执行的任务。例如，一个蛋白质可能是一个转录因子或是一个载体蛋白。生物过程是指基因产物所联系的一个大的生物功能，或者说是它们要完成的一个大的生物目标，如有丝分裂或嘌呤代谢。细胞组分是指一个蛋白质的亚细胞定位，例如一个蛋白可能定位在细胞核中也可能定位在核糖体中。任何一个蛋白质都有可能涉及不只一个分子功能、生物过程或细胞组分中的分类。

基因和基因产物通过注释的过程被划分到不同的 GO 分类中去。每一个给基因添加 GO 注释的作者都必须提供给基因加注释的证据（表 8-7）。图 8-14(a) 所示的就是一个 GO

表 8-7　Gene Ontology 工程中的基因注释证据代码

缩　写	证据代码	例　子
IC	由专家推测得到	一个蛋白质如果有"转录因子"的功能注释，专家就可以推测它定位在细胞核中
IDA	有直接实验证据	如酶活性化验(针对"分子功能")，免疫荧光显微镜(针对"细胞组分")
IEA	有电子注释	基于像 BLAST 这些搜索软件搜索结果的电子注释(与 ISS 比较，IEA 没有被专家证实)
IEP	有表达模式方面的证据	转录水平(如 Nothern blotting 或微阵列实验结果)或蛋白质表达水平(Western blots 实验结果)
IGI	有遗传相互作用方面的证据	抑制基因；遗传致死基因；互补实验；还有其他一些可以提供某一基因功能信息的实验
IMP	由突变表型推测得到	基因突变；基因敲除；过量表达；反义 RNA 实验
IPI	由物理相互作用推测得到	酵母双杂交实验；共纯化实验；免疫共沉淀实验；结合实验
ISS	由序列或结构的相似性推测得到	序列的相似性；结构域；被专家确认的 BLAST 软件的搜索结果
NAS	有报道，但报道信息来源不可知	数据库(如 SwissProt 数据库)登录条目的记录中没有引用已发表的文献
ND	没有相关的生物数据	对应于"unknown"的分子功能、细胞过程或细胞组分
TAS	有报道，且可知报道信息来源	在一些综述文章或字典中可以找到相关信息

注：摘自 http://www.geneontology.org/。

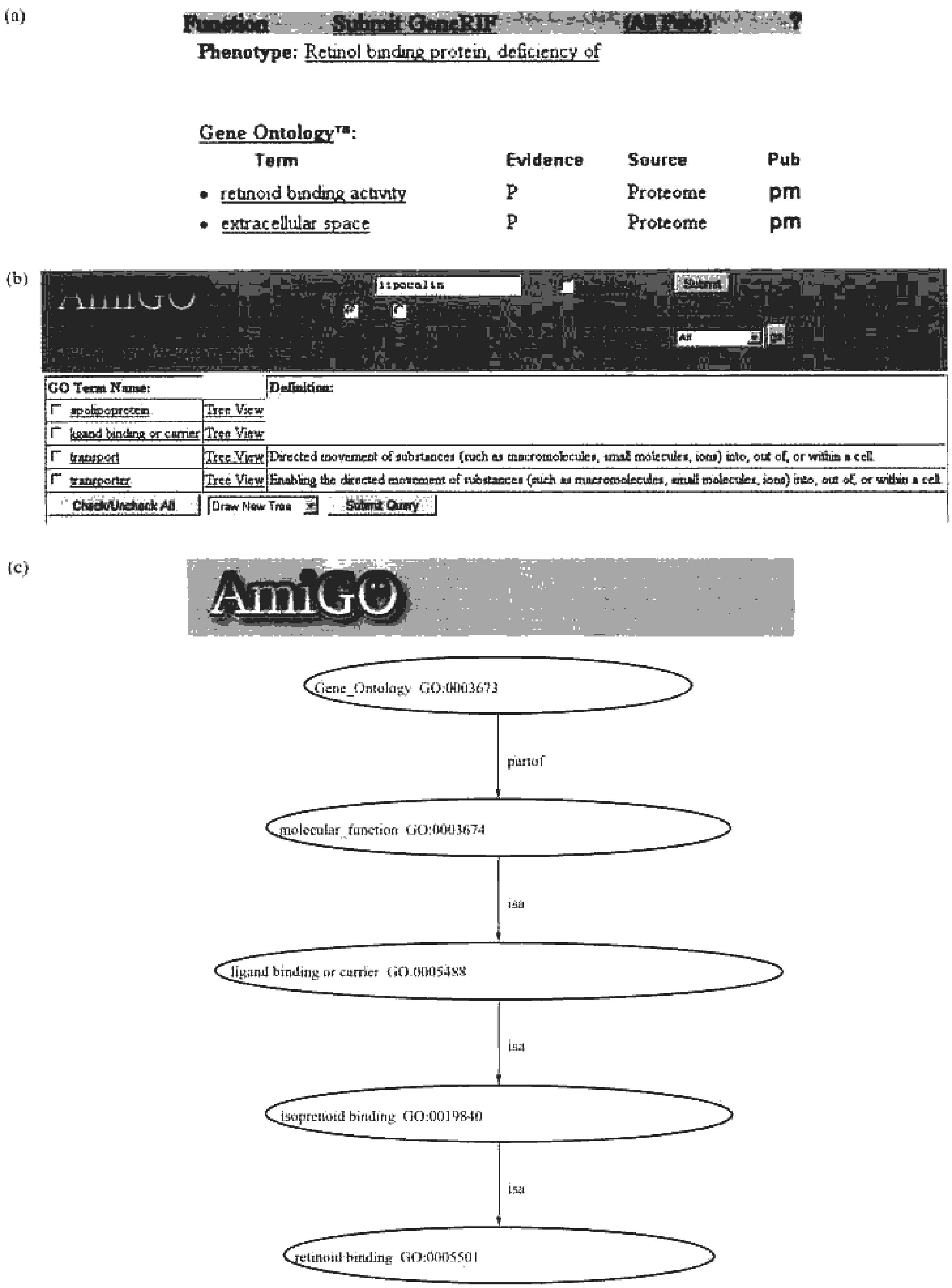


图 8-14 GO协会提供了一组动态而又可控的词汇描述不同物种的基因和基因产物。这组词汇由三个方面组成，它们分别是分子功能、生物过程和细胞组分。(a) 我们可以通过浏览器或地址链接来获取某个基因的GO信息，具体情况可以参见视黄醇结合蛋白（retinol-binding protein）的例子。(b) 点击链接以后会在AmiGO浏览窗口出现GO词汇，包括 (c) 以树状图形式显示的GO词汇之间的关系

注释蛋白的例子，图中所示蛋白是视黄醇结合蛋白（retinol-binding protein）。位置链接登录项（LocusLink entries）包括 OMIM（见第 18 章）、酶学（专门）委员会命名法（见下文）和 GO 词汇等与蛋白质功能相关的信息。对于 RBP4，GO 的注释是“retinoid binding”（分子功能）和“extracellular space”（细胞成分），缺少与生物过程相关的注释。

也可以在 GO 的网页上敲入“lipocalin”关键词来搜索关于 lipocalin 的 GO 注释［图 8-14(b)］。显示的结果中包含一个树状图［图 8-14（c)］，这个树状图表示不同层次 GO 词汇之间的关系。树状图是一个“定向无环图”而不是一个“网络”，这主要是因为每一个子节点（对应于一个 GO 术语）最多只能有一个父节点（也对应于一个 GO 术语）。子节点对应的 GO 术语可能是父节点对应的 GO 术语的一个实例，这种情况下图中父节点到子节点的连线被标记上了“isa”，子节点对应的 GO 术语也可能是父节点对应的 GO 术语的一部分，这时父节点到子节点的连线就会被标记上“partof”。图 8-14(c) 中有“isa”和“partof”关系的例子。

Gene Ontology 协会主站点的网址是 http://www.geneontology.org/。

下面将要考虑蛋白质的定位和功能，这两方面和 GO 分类中的“细胞组分”和“分子功能”有非严格的对应关系。在本章的下文还要讨论蛋白质通路（protein pathways），尽管 GO 分类中的“生物过程”并不是专门与蛋白质通路对应的。

8.2.6 视角 3. 蛋白质定位

细胞定位是一个蛋白质的基本性质之一。mRNA 通过核糖体被翻译成蛋白质。一些蛋白质的翻译是在细胞质中完成的，而另外一些蛋白质（主要是一些将要被分泌到胞外或要被定位到质膜上的蛋白）会被插入到内质网膜上（对于真核生物）或插入到质膜上（对于原核生物）。插入的过程可能是在翻译的同时进行的，也可能是在翻译后进行的，它是由一个单独的识别颗粒（RNA 多蛋白复合体）介导的（Stroud 和 Walter，1999）。在内质网，蛋白质可能会通过第二条途径被转移到高尔基体，然后再进一步定位到细胞内的各种细胞器中（如内含体或溶酶体）或细胞表面。

蛋白质还可能被进一步分泌到细胞外的环境中。将蛋白质运送到合适的地方是通过二级囊泡的转运来实现的。这些囊泡的直径通常有 75～100nm，它们将可溶的或是被限制在膜上的“货物”运输到特定的地方。

也可以根据蛋白质与磷脂双分子层的关系将蛋白质分成两类：一类是可溶的蛋白质，它们存在于胞质、细胞器内腔或胞外环境中；另一类是附着在膜上的蛋白，它们通常与脂双分子层有联系。后一种蛋白有可能被整合到膜内（含有 10～25 个疏水氨基酸残基的跨膜段），也有可能仅仅和膜的表面有联系（通过一些锚定点附着在膜上，见下文）。

很多蛋白质不能被单一地确定存在于细胞一个固定位置上。例如膜联蛋白（annexins）和小 G 蛋白（low-molecular-weight GTP binding proteins）家族就转移于胞质和膜之间（有时在胞质内，有时在膜上）。这种转移运动取决于是否有特定的细胞信号存在，例如钙离子。

蛋白质在细胞中都有合适的定位是因为它们的一级结构（氨基酸残基排列顺序）中固有的细胞定位的信息。例如，如果 KDEL（lys-asp-glu-leu，赖氨酸-天冬氨酸-谷氨酸-亮氨酸）序列出现在一个可溶性蛋白质的羧基末端，那么这个蛋白质将被定位在内质网内。其他类似的模体（motif）也被鉴定出来，这些模体使蛋白质定位到相应的细胞器中，例如线粒体、内吞过程中出现的过氧化物酶体等。只是这些模体并不像 KDEL 模体这样单一（即有些位置上的氨基酸残基可以被别的氨基酸残基替换）。

现在，有很多基于网页的程序可以被用来预测单个蛋白质序列的细胞定位（见8.7网络资源，表8-16）。例如，PSORT程序就准确地预测出了视黄醇结合蛋白（retinol-binding-protein）氨基末端的信号序列（图8-15）。蛋白质序列上的信号肽决定该蛋白质在内质网内部进入哪一条通路再被定位到合适的地方去，目前内质网内部的这些通路还是未解之谜。另一个更普遍的问题是如何确定蛋白质中的跨膜区域。诸如TMPred这样的程序（见8.7网络资源，表8-17）可以解决这个问题（图8-16）。但是就像前面所说的，在利用一些算法预测蛋白质性质（如跨膜区拓扑结构）时需要注意，因为预测的结果可能不准确。

Results of Subprograms

```
PSG:  a new signal peptide prediction method
      N-region:  length 2;  pos.chg 1;  neg.chg 0
      H-region:  length 14;  peak value  10.03
      PSG score:    5.63

GvH:  von Heijne's method for signal seq. recognition
      GvH score (threshold: -2.1):    3.93
      possible cleavage site: between 16 and 17

>>> Seems to have a cleavable signal peptide (1 to 16)
```

图8-15 PSORT服务器（http://psort.nibb.ac.jp/）提供了一个以网页为基础的预测蛋白质亚细胞定位的服务。这个程序会搜索决定蛋白质细胞定位的分类信号。以视黄醇结合蛋白为例，它的输出结果显示了有很强的证据说明这个蛋白质有信号肽并且信号肽断裂的位置在第16～第17个氨基酸残基之间。这种信号肽标志着这个蛋白质会进入分泌通路并且有一些蛋白（如RBP）将会被分泌到胞外

Michael Snyder和他的同事第一次尝试着在整个蛋白质组范围内对蛋白质的细胞定位进行分析（Kumar等，2002）。他们克隆了芽殖酵母（budding yeast *S. cerevisiae*）中数千个cDNA序列，这些cDNA序列编码了数以千计的蛋白质，然后在cDNA的3′端或5′端连上抗原表位标签，用这种方法将抗原表位标签重组到酵母的基因组中去。这种定向重组方法可以让系统地同时观察多个感兴趣的蛋白在细胞中的定位。作为补充，他们还用了随机转座子介导的标记基因的方法。

Kumar等（2002）一共重组获得了超过13000个酵母株，然后用相应的带有荧光物质的单克隆抗体去与特定的抗原表位结合，利用免疫荧光显微镜的方法确定了2744个酵母蛋白的细胞定位。很多功能不清楚的蛋白通过这种方法确定了它们在细胞中的定位，从而让我们对它们的功能有了一定的了解。例如，如果一个蛋白质被定位在过氧化物酶体中，那么它的功能就可能是参与脂肪代谢。

真核细胞中，位于细胞内的细胞器的膜占整个细胞所有膜比例的95%。

抗原表位标签是一个短的蛋白质片段，如9个氨基酸残基组成的红细胞凝集素（HA）肽段，它可以共价地结合在我们感兴趣的蛋白质上，然后用特定的抗血清检测抗原表位标签，结合免疫荧光显微镜或是免疫共沉淀就可以确定我们感兴趣的蛋白质在细胞中的定位。

细菌的转座子是可移动的DNA元件，它可以随机插入到基因组DNA序列中。转座子可以和一个抗原表位标签合并。这种策略在实际中比直接克隆酵母中特定的cDNA要简单，尽管从实验的角度看，用转座子标记的蛋白质可能在细胞中很难定位。

收录了Kumar等（2002）实验所得的2900张荧光显微照片的数据库的网址是http://ygac.med.yele.edu。

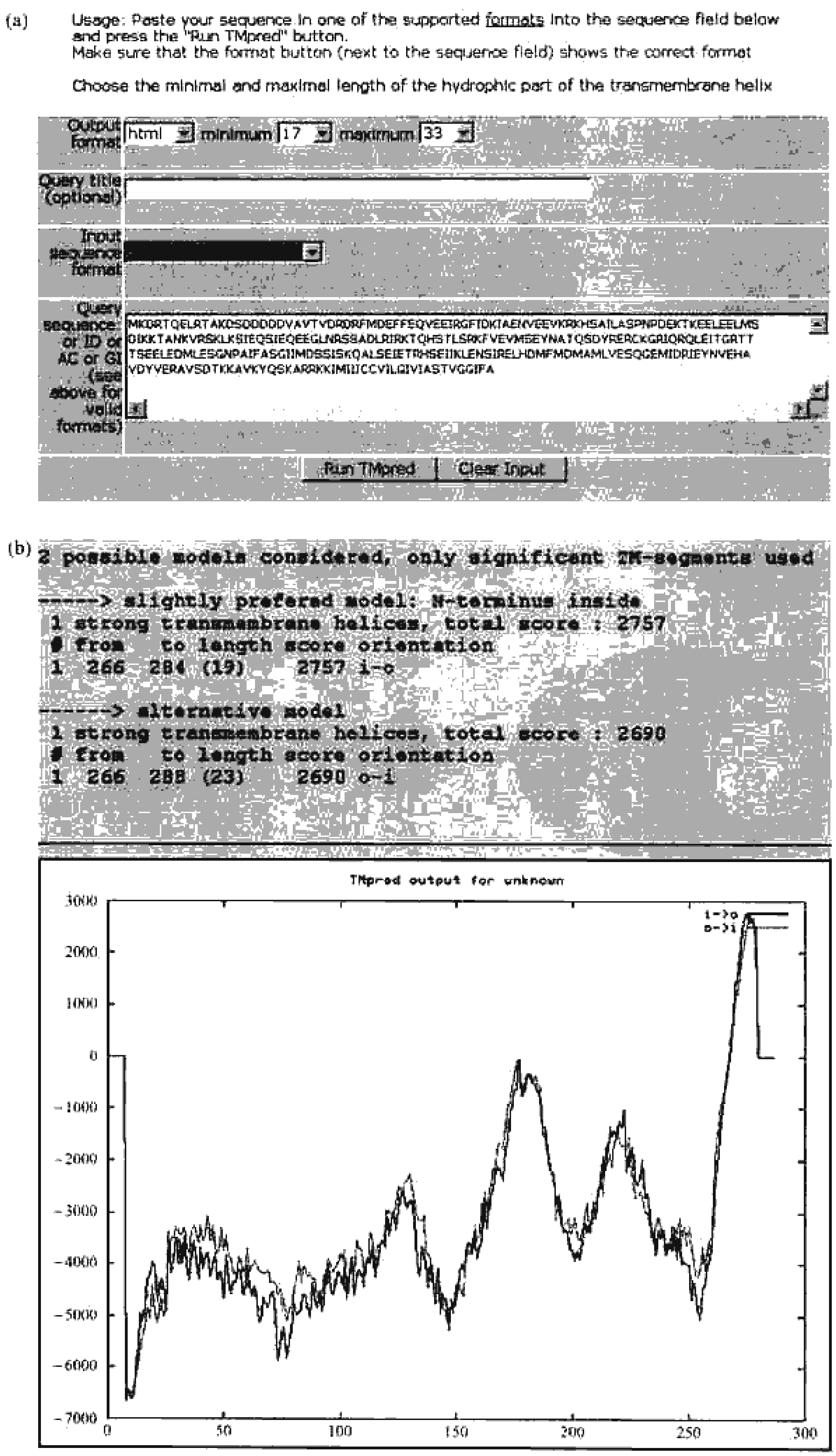


图 8-16 有很多软件可以用来预测蛋白质的跨膜区。(a) 将突触融合蛋白 (syntaxin) 的序列作为 TMpred 程序 (http://www.ch.embnet.org/software/TMPRED_form.html) 的输入。(b) 程序的输出结果显示预测出来的模型是在蛋白质的羧基末端有一个单独的跨膜区结构域 (已被实验所证实)

8.2.7 视角 4. 蛋白质的功能

前面已经讨论了生物信息学工具和方法在蛋白质家族、蛋白质物理性质和蛋白质细胞定位方面的应用。我们要讨论的第四个方面就是蛋白质的功能。蛋白质功能的定义是它们在细胞中所起的作用 (Jacq, 2001)。每一个蛋白质都是基因的产物，它们以不同的方式与细胞

环境接触以促进细胞的生长和行使它们的功能。我们可以从如下几个方面来考虑功能这个概念（图 8-17）。

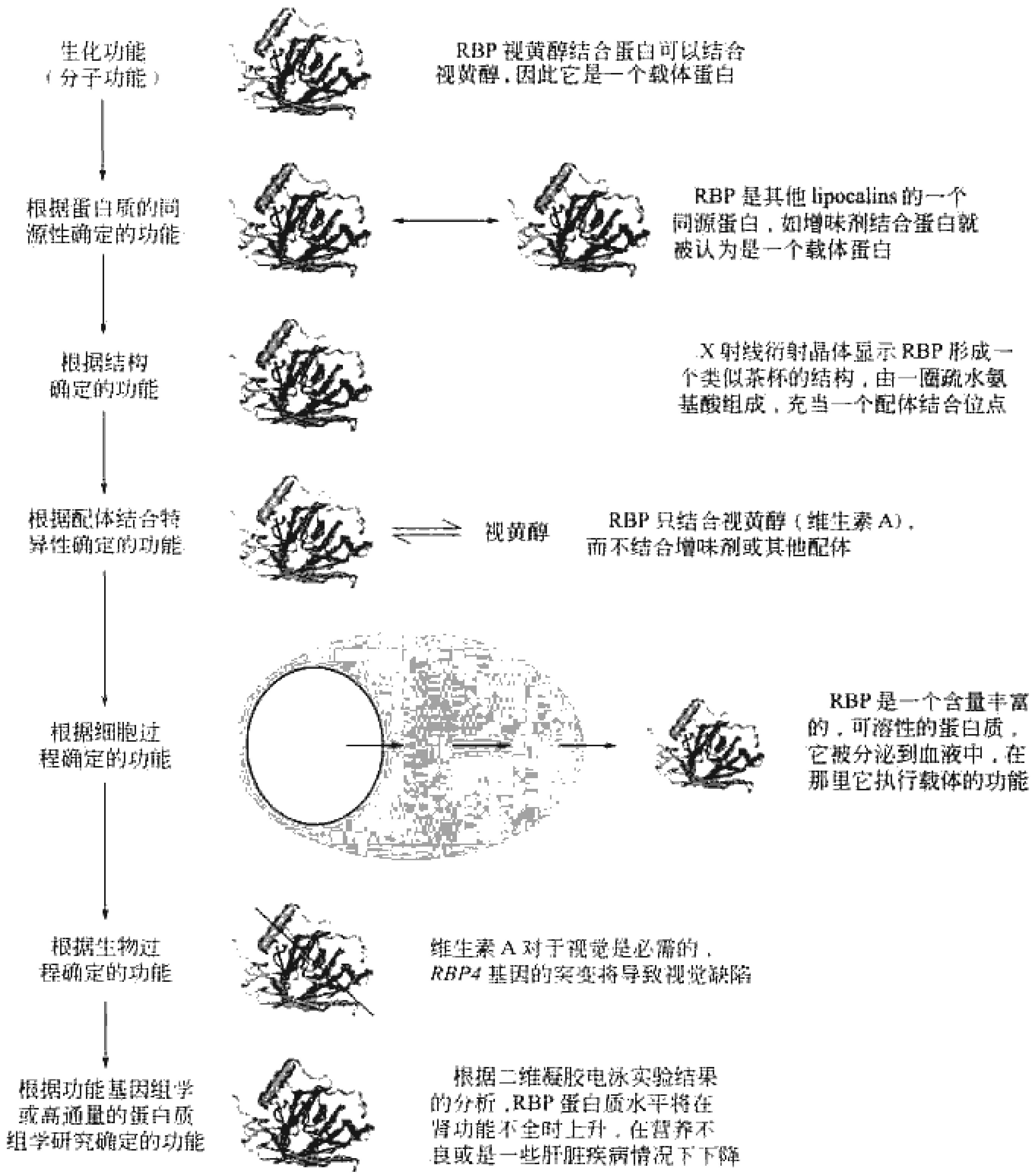


图 8-17 可以从多个角度去研究蛋白质的功能，视黄醇结合蛋白就是一个例子

① 每个蛋白质都有生化功能，生化功能和蛋白质的分子功能的内在含义是一致的。对于酶而言，它的生化功能就是催化一种或几种底物转化为其他物质。对于一个结构蛋白而言，它的生化功能就是影响细胞的形状。对于一个转运蛋白而言，它的生化功能就是将一个配体由一个地方转移到另一个地方［这种转运在不需要能量物质（如 ATP）的情况下也可能发生，如视黄醇结合蛋白在血浆中可以转运视黄醇］。对于那些从基因组序列中预测出来的尚未被实验证实的蛋白质，它的生化功能是未知的，但是功能一定存在。人们认为细胞中不存在没有任何功能的蛋白。

② 通常利用搜索同源蛋白的方法预测一个蛋白质的功能。如果一个预测出来的蛋白质有一个同源蛋白，且同源蛋白是一个酶，那么预测出的蛋白质也会被注释上具有酶的功能。对用这种方法确定的功能，我们最好把它当成一个假设，因为它还需要进一步的实验验证。

③ 也可以在结构的基础上预测一个蛋白质的功能（见第9章）。如果一个蛋白的三维结构与一个功能已知的蛋白的三维结构采取了相似的折叠方式的话，我们就有一定的根据说它们的功能也类似。但有一点需要注意，结构上的相似并不代表它们是同源的，同源也不等价于功能相同。

④ 所有蛋白质都是在其他蛋白质或分子存在的环境下行使功能的，因此，一个蛋白质功能的定义应该包括它们的配体（如果该蛋白是一个受体蛋白）、底物（如果该蛋白是一个酶）和它接触的脂质（如果该蛋白与膜有联系的话），或是其他与该蛋白有接触的分子。增味剂结合蛋白（odorant-binding proteins，OBP）是 lipocin 蛋白家族中的一个成员，可以在鼻腔黏液中结合很多种类的增味剂，这说明这种蛋白和其他物质结合的性质是它的重要功能（Pevsner 等，1990）。但是如果只从 OBP 与其他物质结合的性质出发，将无法真正解开 OBP 功能的面纱。事实上，OBP 还可以将增味剂转运到嗅觉上皮细胞来促进嗅觉感知，它还可以将增味剂从嗅觉上皮细胞上清除，或者说它还参与代谢增味剂的代谢。

⑤ 很多蛋白质只在一个生化路径的某一步中起作用，例如三羧酸循环，这个循环的每一步都需要一些蛋白来执行一个非常复杂的任务。其他的例子还有过氧化物酶体中的脂肪酸氧化和蛋白酶体中的蛋白质降解等。

⑥ 很多蛋白质执行功能的过程充当了一个很大的生物过程的一部分。这些生物过程包括细胞分裂、生长和衰老；神经元细胞的轴突要派生出分支，向前伸展，识别目标和形成突触；所有的细胞都需要通过不同的途径向胞外分泌分子等。所有的细胞过程都要求蛋白质有条不紊地按一定顺序来执行它们的功能，单个蛋白质的功能需要在它所参与的整个大的功能的背景上进行定义。

⑦ 我们还可以在整个基因组编码的所有蛋白质——蛋白质组水平上来考虑一个蛋白质的功能。功能基因组学这个术语是指利用实验的手段或计算工具对成百上千个表达的基因进行探索性的研究和分析。因为大多数转录的最终产物还是蛋白质，所以功能基因组学这个词语也经常指大规模地对蛋白质功能进行研究。

从上文可以看出，对蛋白质的定义有多种方法。很多蛋白质都属于酶类，酶学委员会（Enzyme Commission，EC）根据一个标准的命名规则给 4000 多个酶（表 8-8）提供了命名。当一个基因组被测序后，那些有潜在可能是编码蛋白质的序列被找了出来，如果这些序列中有与 EC 提供的列表中的酶同源的序列，这些找到了同源酶的潜在基因的功能就有了特定的可验证的假设——我们可以假设它们与同源酶具有相同或相似的功能。

表 8-8 根据酶活性对蛋白质功能进行的划分：EC 分类系统的一部分列表（27.0 版本，2001年10月）

EC号	类别描述	酶的数目	子类的例子
1.－.－.－	氧化还原酶	1003	
1.1.－.－	—	—	作用于 CH—OH 基团
1.2.－.－		—	作用于醛类或是氧络基团
2.－.－.－	转移酶	1076	
2.1.－.－	—	—	转移一碳基团
3.－.－.－	水解酶	1125	
4.－.－.－	裂解酶	356	
5.－.－.－	异构酶	156	
6.－.－.－	连接酶	126	

注：摘自 http://kr.expasy.org/cgi-bin/enzyme-search-cl。

另外一种使用范围更广的确定蛋白质功能的方法就是利用COG（Clusters of Orthologous Groups）数据库（见第14章）。表8-9列出了这个数据库系统定义的功能分类。虽然COG最初关注的是原核生物基因组，但是基本的分类对所有物种的细胞过程适用。不过，很多真核生物特有的功能（如细胞凋亡和复杂的发育过程）在COG中是没有的。

表8-9 COG（Clusters of Orthologous Groups）数据库中蛋白质的功能分类

主要分类	功能	直系同源组的数目	结构域
信息存储和处理	翻译、核糖体结构和生物合成	217	6449
	转录	133	5442
	DNA复制、重组和修复	184	5337
细胞过程	细胞分裂和染色体分裂	32	842
	翻译后修饰，蛋白质翻转，分子伴侣	109	3155
	细胞包裹物质的合成，外围膜	155	4079
	细胞运动和分泌	133	3110
	无机离子转运和代谢	160	5112
	信号转导	96	3623
代谢	能量产生和转化	223	5584
	碳水化合物的转运和代谢	170	5257
	氨基酸的转运和代谢	233	8383
	核苷酸的转运和代谢	85	2364
	辅酶的代谢	154	4057
	脂类的代谢	75	2609
	二级代谢物的合成、转运和分解代谢	62	2754
描述信息很少	仅有简单的预测功能	449	11948
	未知功能	752	6431

注：摘自 http://www.ncbi.nlm.nih.gov/COG/（2002年9月）。

Gene Ontology 协会（Ashburner 等，2000）对生物过程（biological process）的定义是基因或基因产物所贡献的生物目标，一个生物过程通过一个或多个有序的分子功能组合来完成。生物过程通常伴随物质的物理和化学转化，从这个意义上说一些物质进入一个生物过程后，等它们再出来时就已经不是原来的物质了。

细胞凋亡是细胞的程序性死亡，这种现象在很多多细胞生物中都会出现，既可能作为一个正常的过程在发育过程出现，也可能作为组织自我稳定的机制在成熟的器官中发生。细胞凋亡过程有可能会由外在的刺激（如一些化学试剂或毒素）触发，也可能会由体内的物质（如引起氧化性损伤的物质）触发。

8.3 蛋白质组学：对高通量蛋白质数据进行分析的生物信息学工具和方法

传统的分析蛋白质的方法是对单个蛋白质的功能进行实验分析（如酶活性以及它对细胞过程的影响）。这些实验分析是建立在蛋白质纯化技术上的，分析之前必须将要研究的蛋白质从它所在的环境中提取出来。目前，我们已经通过这种方法对数千种蛋白质进行了单独的分析研究。每个蛋白的生化性质都不相同，它们因为自身大小、所带电荷或疏水性等造成的与各种树脂的亲和力也不相同。

除了能单独研究一个蛋白质外，同时对数千个蛋白进行高通量的分析现在也已成为了可能（Molloy和Witzmann，2002）。我们将要介绍4种对蛋白质进行高通量分析的手段，它

们分别是二维凝胶电泳、亲和层析串联质谱、酵母双杂交系统和基于基因组 DNA 序列分析的被称为“Rosetta Stone”的计算方法。图 8-18 展示了这些方法。

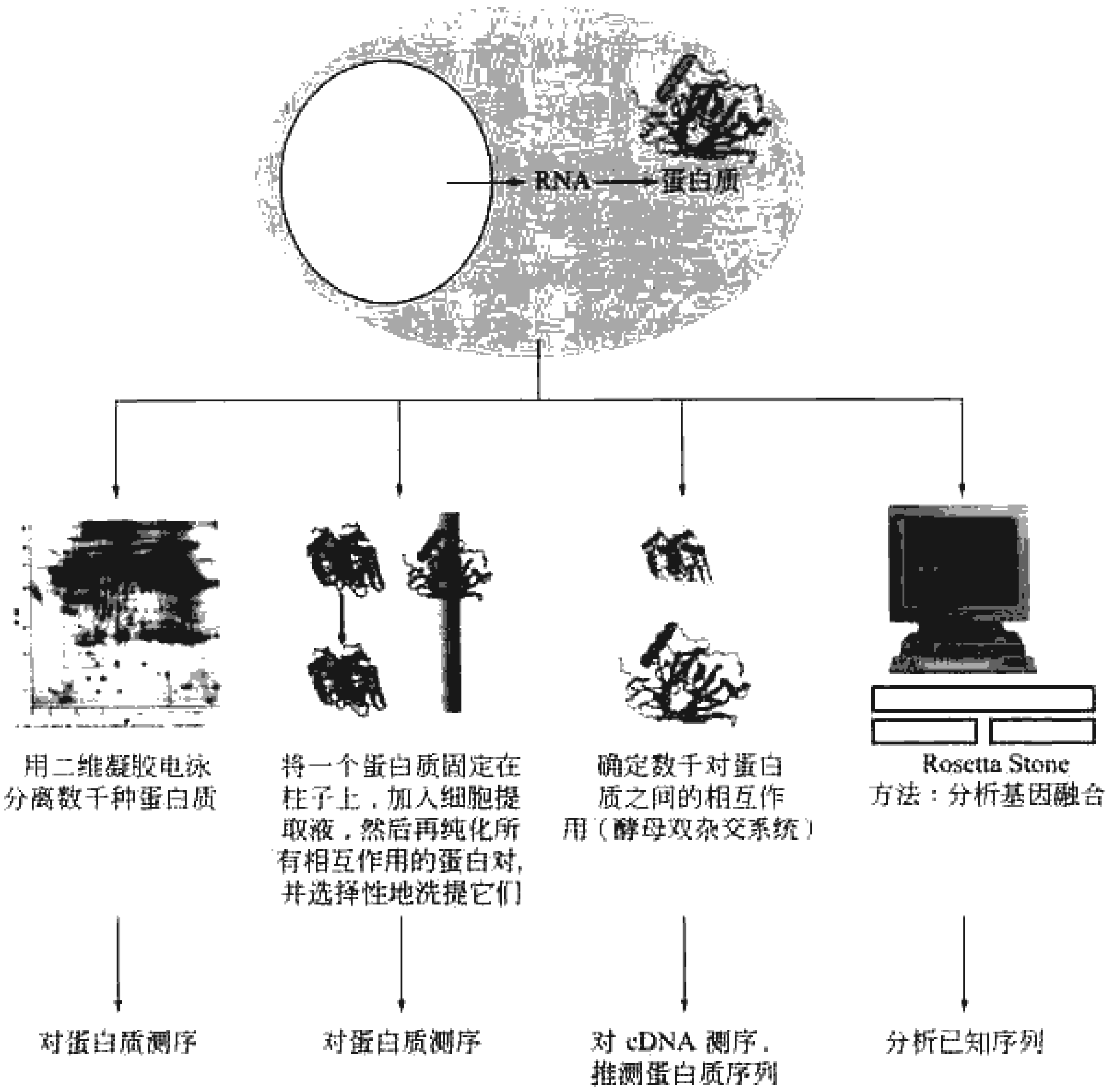


图 8-18　进行高通量蛋白质分析的方法。其他的分析策略（如蛋白质芯片，含有固定蛋白质的高密度微阵列）目前也在开发之中

大多数蛋白质的功能都是未知的。即使对于像大肠杆菌（*Escherichia coli*）和芽殖酵母（*S. cerevisiae*）这样相对研究较多的模式生物，也仅仅是一半蛋白质被指定了功能。高通量的蛋白质组工程能够大范围地为蛋白质指定功能，能确定蛋白质在哪种特殊的生理条件下会出现，还能确定哪些蛋白质之间有相互作用。

还有一些其他高通量的实验方法用于蛋白质组分析，如蛋白质芯片，它类似于 DNA 芯片，是由一些附着在固体材料上具有亲和力的反应试剂（如特定的抗体）构成的（MacBeach，2002）。这种技术还没有被大规模地运用是因为很难保证被固定的蛋白质的结构和功能不发生变化。组织芯片是另外一种高通量的实验技术，它特别适合分子病理学的研究（Kononen 等，1998；Kalliioniemi 等，2001）。组织芯片是将几百个或几千个组织标本按照一定的顺序固定在一个固体材料上，利用这些技术可以并行检测样品中 DNA、RNA 和靶蛋白质的芯片。

理解单个生物中基因和蛋白质功能最广泛的方法就是研究芽殖酵母，我们将在第 15 章中详细讨论研究芽殖酵母的方法。

8.3.1　二维凝胶电泳

聚丙烯酰胺凝胶电泳的方法是对蛋白质分子量进行分析的最常用和最主要的手段。蛋白质（或核酸）因为带电荷，所以将它们放到电场中时它们会做迁移运动。变性的蛋白质通过惰化（这样它就不会与蛋白质结合）和多空化（这样蛋白质才能通过）处理的丙烯酰胺凝胶

的网格时，它的迁移速率和它的大小近似成反比关系。因此一次电泳实验就可以将含有多种蛋白质的混合液的组分彼此分开。为了让蛋白质变性，电泳环境中需加入变性剂十二烷基硫酸钠（SDS），所以这个技术通常被简称为 SDS-PAGE。

O'Farrell（1975）根据蛋白质按照所带电荷的不同可以分开这个事实将电泳技术扩展到了二维。在其中一维上进行等电聚焦使蛋白质根据电荷的不同初步分离；另一维上依旧利用凝胶网格对每一个连续的 pH 梯度上的蛋白质都进行凝胶电泳。pH 梯度的范围通常是 3～11。每一个蛋白质都是一个两性离子，在进行二维电泳的时候，它们会自动迁移到使自身的净电荷为 0 的 pH 梯度上。这个 pH 值就是这个蛋白值的等电点（简称为 pI）。一个复杂的蛋白质集合体能根据它们所带电荷性质的不同（对应于第一维等电聚焦）和分子大小的不同（对应于第二维 SDS-PAGE）而分开。现在，二维凝胶电泳已经成为一个对蛋白质组进行分析的成熟方法了（Dunn，2000）。

图 8-19 所示的是一个二维凝胶电泳实验结果的例子。从人的脑脊髓液中得到数百微克的蛋白质混合液在二维凝胶电泳中按照蛋白质的 pI（x 轴上的等电聚焦）和分子大小（y 轴上的 SDS-PAGE）的不同被分离开了。再利用可以和蛋白质结合的染料（例如硝酸银或考马斯亮蓝）我们就可以直接观察数千种蛋白质了。需要注意的是，有些蛋白质的含量很高。很多蛋白质形成的点在第一维上有特征分布。第一维上点的分布被人们称为“charge train”，“charge train”经常表示一系列蛋白质，它们含有不同数量的同一种带电基团，如共价结合的磷酸基团。

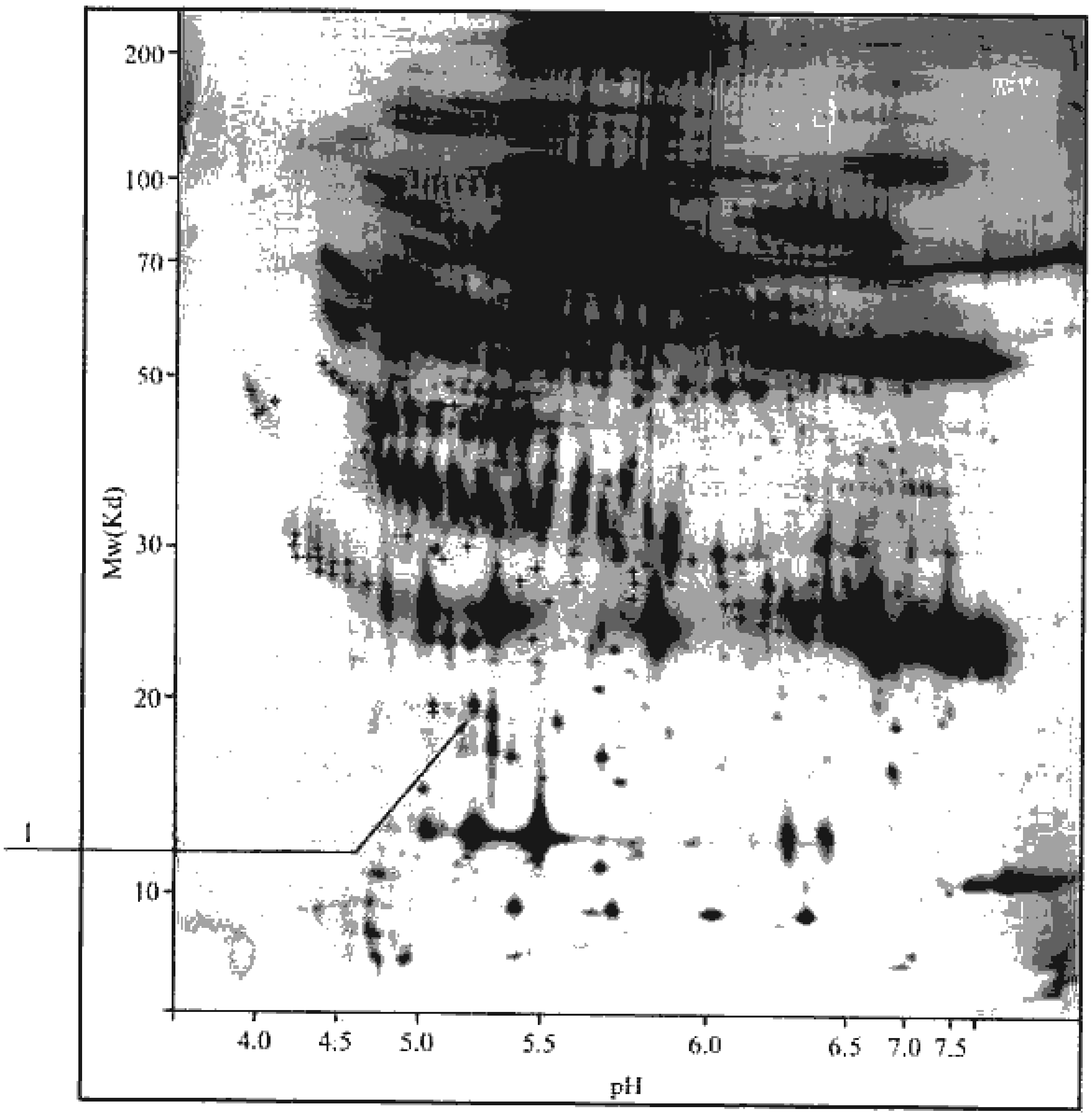
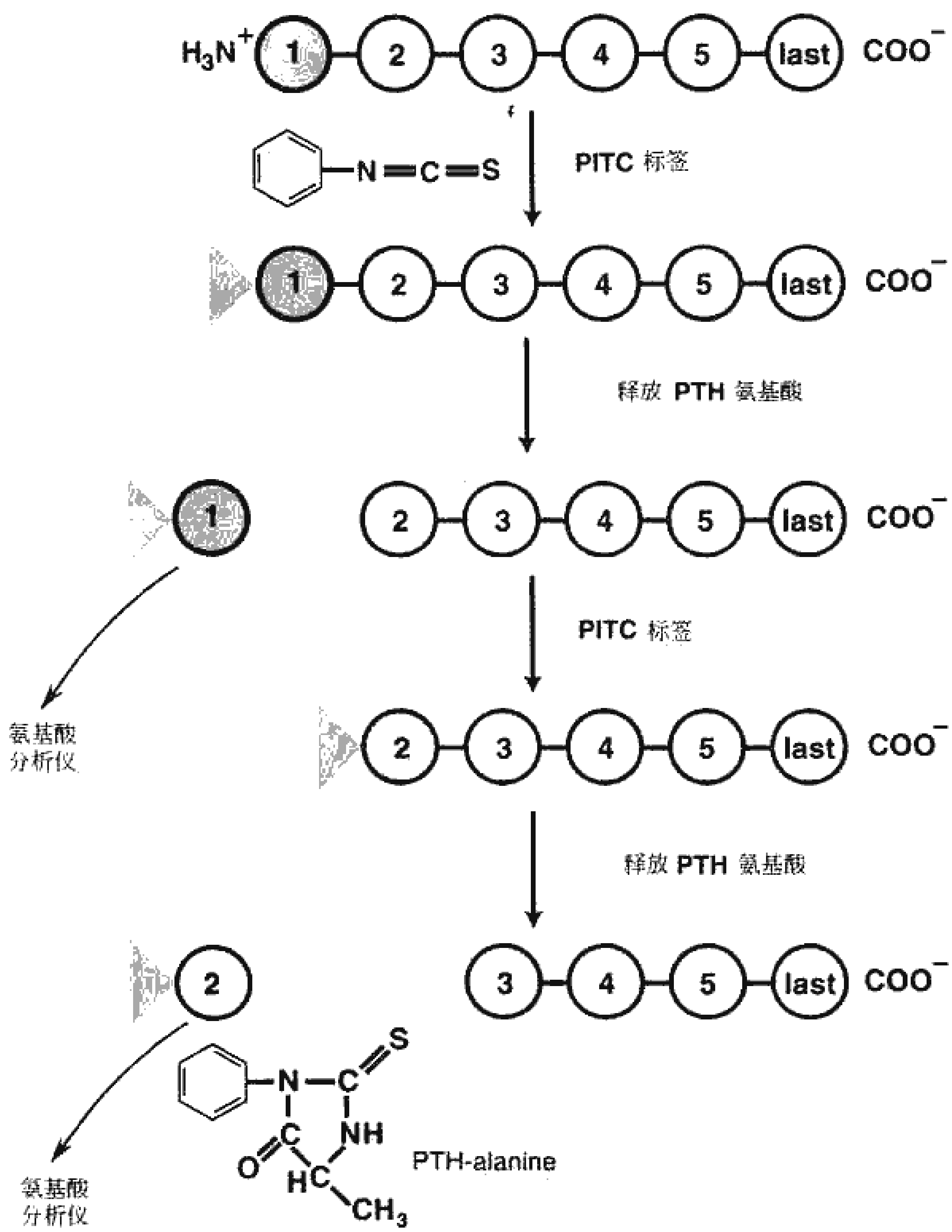


图 8-19 ExPASy 数据库收录的脑脊液的二维凝胶图谱。每一条蛋白质带对应的蛋白质都已用微量测序仪确定了蛋白质序列，点击每条带可以看到对相应蛋白质的描述。箭头 1 指示的是视黄醇结合蛋白对应的条带，它的等电点约为 5.2，分子量约为 20000

二维凝胶电泳一个重要的性质就是被分离出来的组分可以通过直接的蛋白质测序（框 8-1）或是比较分辨率较高的质谱技术（如基质辅助激光解析-电离飞行时间质谱，matrix-assisted laser desorption/ionization time-of-flight spectroscopy，简称 MALDI-TOF）（Farmer 和 Caprioli，1998）

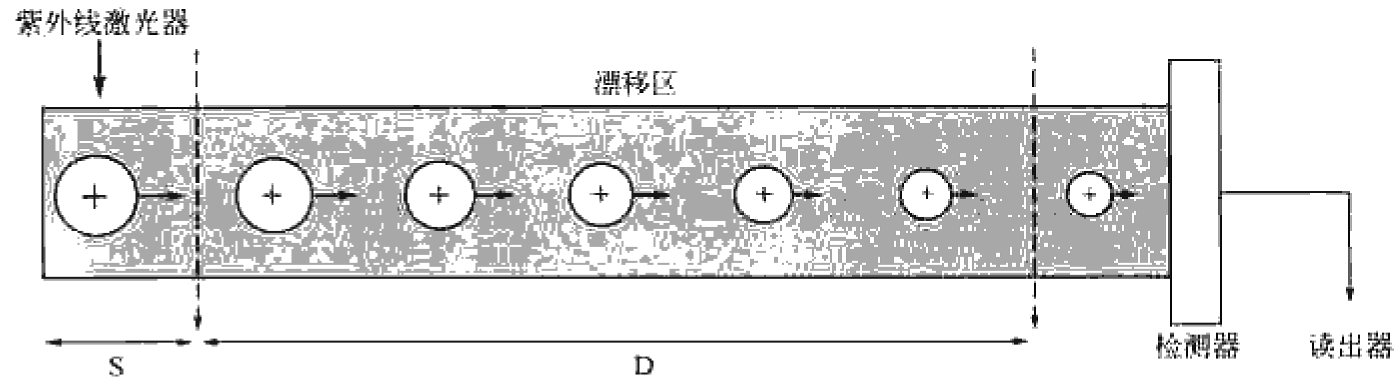
框 8-1 Edman 降解法测蛋白质的序列

将氨基酸片段转移到聚偏二氟乙烯膜（membrane polyvinylidene fluoride，简称 PVDF）上，然后将附有氨基酸片段的膜放置在微量测序的仪器中，这个仪器将对氨基酸片段进行连续的 Edman 降解。在 80%的时间里，氨基酸片段的氨基末端是被保护起来的（如通过酰化或是让氨基末端不可接近的方法）。一个标准的过程是首先用蛋白酶（如胰蛋白酶）将蛋白质水解，然后用反相高压液相色谱来纯化水解后的氨基酸片段，再用质谱（MALDI-TOF 质谱）检测纯化后片段的纯度和分子量，最后再进行 Edman 降解过程。下面通过一个 6 个氨基酸残基构成的片段为例来说明 Edman 降解的过程。氨基末端的第一个氨基酸残基与异硫氰酸苯酯（PITC）偶联，这会导致这个氨基酸残基环化并从蛋白质链上脱落下来，最终形成 PTH 氨基酸。仪器会自动分析 PTH 氨基酸的类型。这个循环会重复进行，直到每一个氨基酸残基都被鉴定出来为止。以 PTH 丙氨酸为例，它的结构在下图中被显示出来。一般的结果是读出 10～20 个氨基酸残基的序列。其对应的蛋白质或基因可以通过 BLAST 软件进行搜索。



框 8-2 基质辅助激光解析-电离飞行时间质谱（matrix-assisted laser desorption/ionization time-of-flight spectroscopy，简称 MALDI-TOF）

质谱是一项确定蛋白质样品或其他大分子分子量的技术。样品被置于一种可以吸收紫外线的基质中。激光在源区击中样品。在基质的环境中样品被离子化，其中一些可能会消失（如可能会发生吸收）。在电场存在的情况下，会出现电离现象，加速的离子进入一个长的漂流区（D）。每个蛋白质片段的加速度将和离子的质量成正比。



（见框 8-2）来确定它们到底是哪一种蛋白质。二维 SDS-PAGE 的应用包括对在人和大鼠脑组织中表达的蛋白进行观察（Langen 等，1999，Fountoulakis 等，1999）和对膀胱肿瘤组织中异常表达蛋白的分析（stergaard 等，1997）。Grunenfelder 等（2001）利用二维凝胶电泳研究了一种细菌（*Caulobacter crecentus*）在整个细胞周期中蛋白质的合成，他检测到的蛋白质占所有预测的基因产物（979）的 25%。其中很多蛋白质在一个单独的周期内就降解了。如果将二维凝胶电泳的实验结果和利用微阵列得到的 mRNA 表达谱结合起来会是一件很有意义的事情。

ExPASy 是提供二维凝胶电泳实验结果信息的主要数据库（Hoogland 等，1999；Sanchez 等，2001）。各个物种和各种实验条件下的实验结果都可以在这个数据库中得到。通过点击数据库网站上的凝胶图标（图 8-20），我们可以选择查看哪个实验结果。网页上的每个凝胶图标都链接

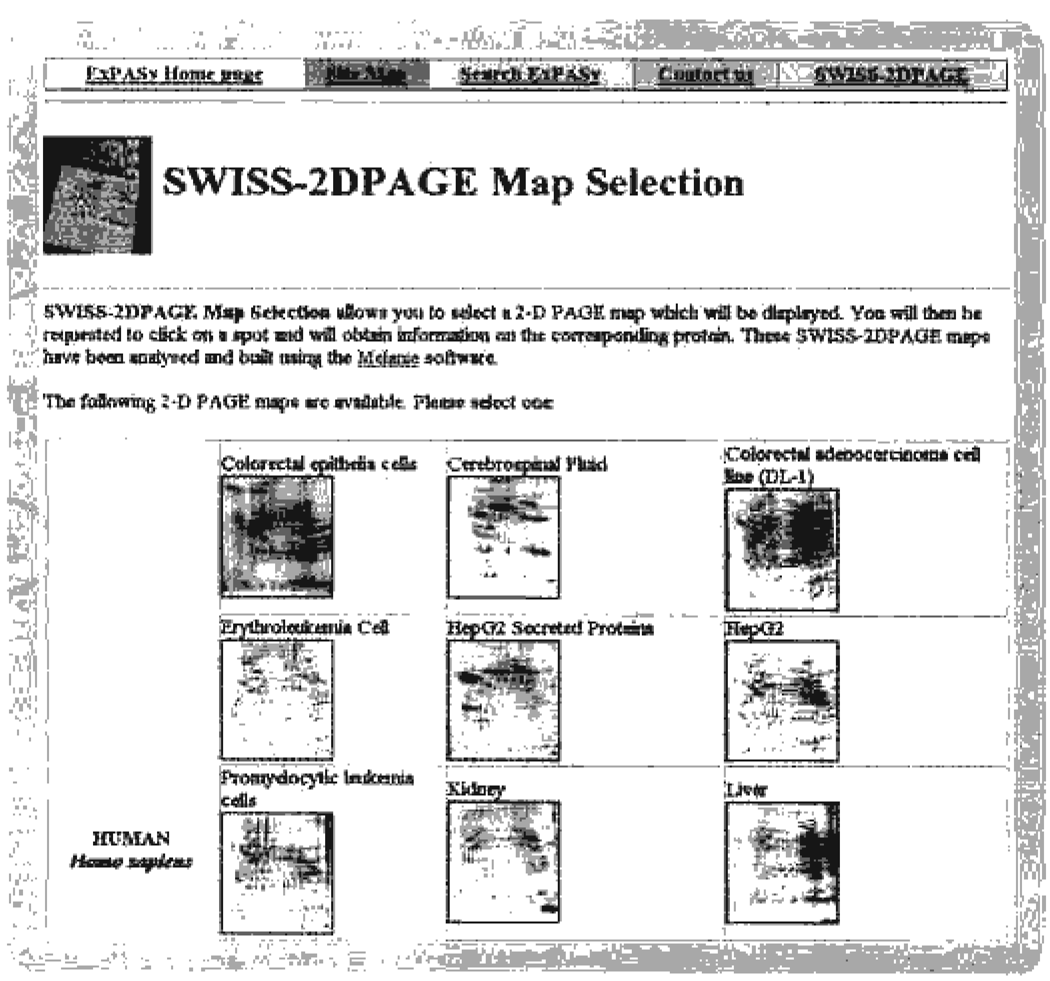


图 8-20 ExPASy 的 Swiss-2DPAGE（二维聚丙烯酰胺凝胶电泳）站点提供了很多物种中的蛋白质的二维凝胶电泳图谱，这些物种包括人类（上面已经显示）、小鼠、植物和细菌，点击每一个凝胶图谱可以查看更为详细的信息

到一个类似于图 8-19 所示的图上。也可以通过别的方式在数据库中进行搜索，比如关键字。当输入“retinol”（视黄醇）作为搜索词时，将会出现 4 个二维凝胶实验的结果（图 8-21），在这 4 个实验中都有视黄醇结合蛋白（retinol-binding protein）被鉴定出来了。进一步选择 4 个实验结果中的一个，将会出现如图 8-22 所示的图，图中我们感兴趣的视黄醇结合蛋白被突出显示出来。

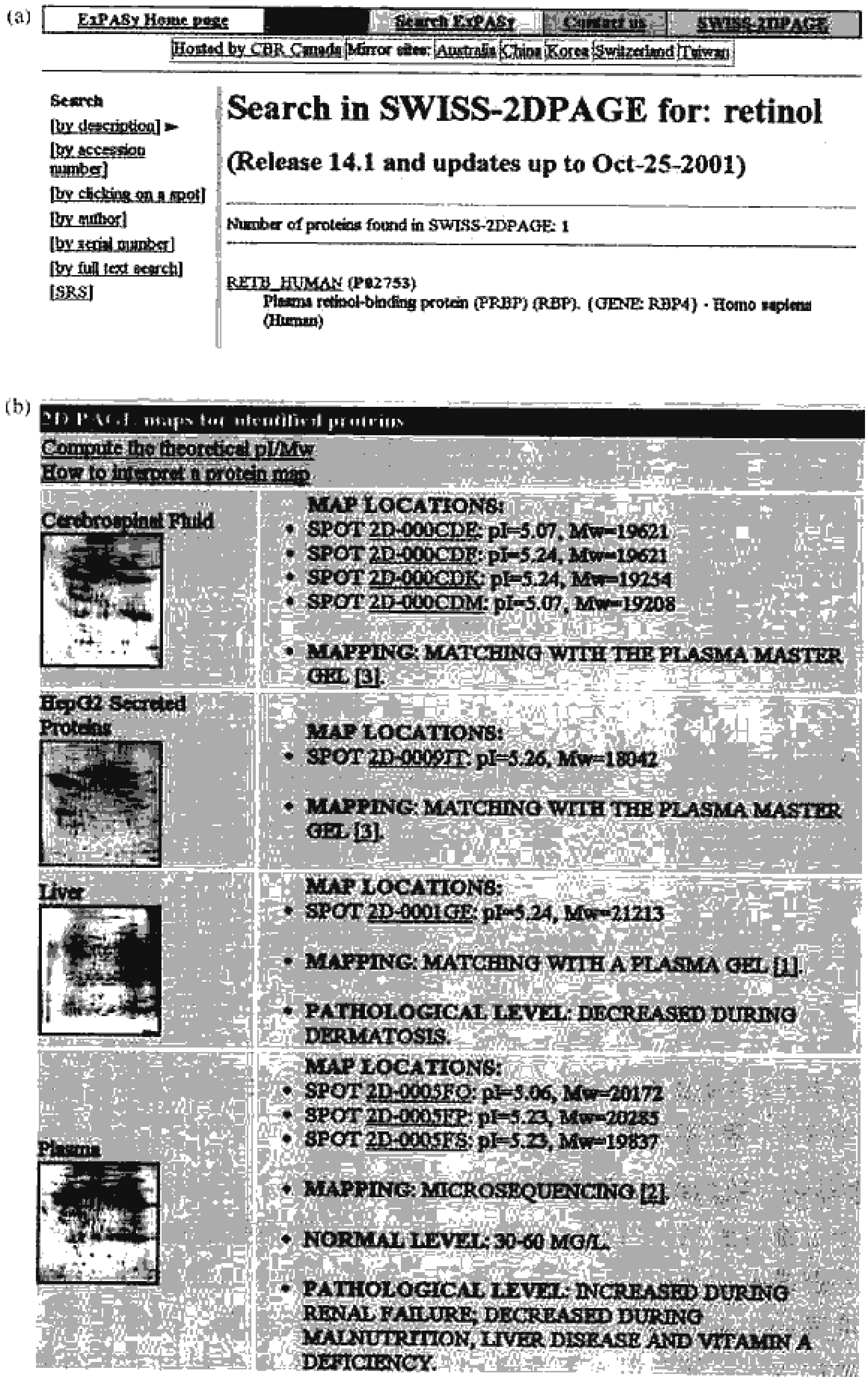


图 8-21 Swiss-2DPAGE 站点可以通过关键词进行检索。(a) 用关键词“retinol”进行搜索，结果页面 (b) 显示了一系列链接到那些已知含有视黄醇结合蛋白（RBP）的二维凝胶图谱上。注意除了链接以外，站点还提供了一些附加信息：在血浆中，检测出的 RBP 的不同水平与 RBP 参与的不同通路状态有关。另外需要注意的就是脑脊液中 RBP 在二维凝胶图谱的四个不同的位置上被检测出来，这四个位置上的 RBP 等电点和分子量都略有差别。二维凝胶图谱上位置的不同通常反映实验状态的差异和蛋白质翻译后修饰的不同

二维凝胶电泳（2DG）方法有如下几点限制。

① 它不能对多个样品进行并行的高通量处理。

② 经常是只有那些在样品中含量较高的蛋白质才能被检测到。疏水性蛋白质（包括含有跨膜区结构域的蛋白质）不能在二维凝胶中显示出来。

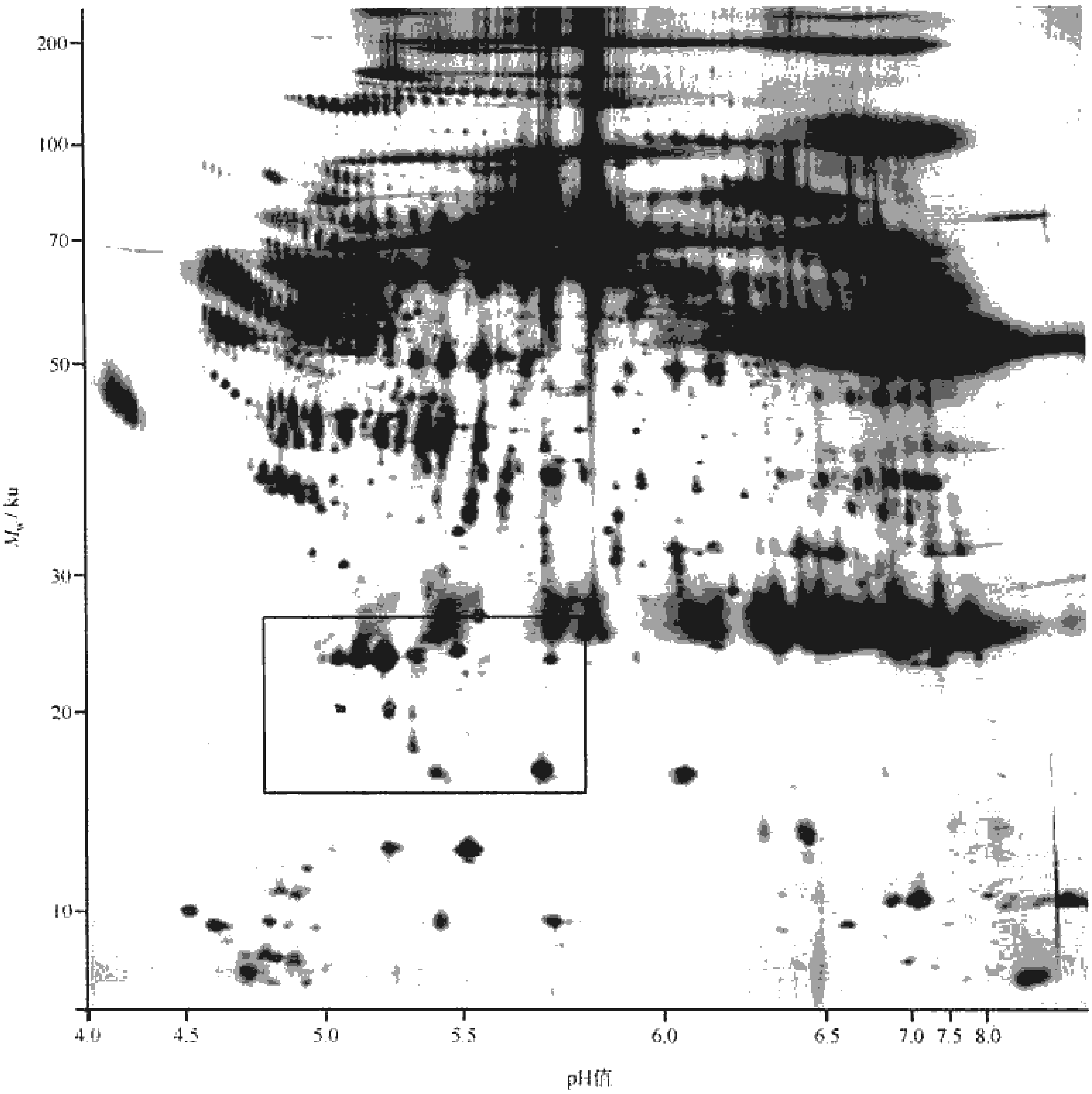


图 8-22 人类脑脊液（CSF）的二维凝胶图谱上的 RBP 的位置。这张图谱可以通过点击 CSF 小图标获得［图 8-21(b)］。注意这张图谱与图 8-19 显示的图谱很相似（但还是有区别）。这种差别经常出现，是由进行二维凝胶电泳的实验条件不同而造成的（如跑胶时间的长短、缓冲液的浓度、两性化合物的浓度、凝胶中聚丙烯酰胺的百分含量和蛋白质滤过的方法）

我们在第 6 章提出过 mRNA-蛋白质的关联问题。

③ 要通过二维凝胶电泳得到一致的结果需要掌握很强的专业技术。在比较相同样品的两个二维凝胶谱（2DG profiles）时，如果聚丙烯酰胺凝胶有略微的差别或电泳是在不同状态下进行，很难对两个二维凝胶谱上的点做准确的比对。为了能让二维凝胶电泳实验可重复，技术上有一个重要的改进，就是引入不变的 pH 梯度。这个操作利用了干的覆盖物。第二次实验时将前一次实验的 pH 梯度替换掉。

8.3.2 亲和层析和质谱

亲和层析这个技术中，配体（如蛋白质）会结合在柱子的网格上，这种结合是化学性

的结合。目前，世界上有两个研究小组致力于研究确定芽殖酵母（*S. cerevisiae*）中存在的数千种多蛋白复合物的方法（Gavin 等，2002；Ho 等，2002；reviewed in Kumar and Snyder，2002）。每个小组都挑选了大量的带有标记的“诱饵”（bait）蛋白，然后这些带有标记的“诱饵”蛋白的基因被导入酵母中，在那里它们将形成酵母本身的蛋白复合物。在与生理相关的条件下，蛋白复合物形成后，再将带有标记的“诱饵”蛋白提取并纯化出来，同时提取和纯化出来的还有与这些“诱饵”蛋白结合的蛋白质。然后利用一维 SDS-PAGE 将得到的蛋白复合物中的蛋白质分开。用剃刀将胶上（对于不同的“诱饵”蛋白有不同的胶）的蛋白质带切下来，用胰岛素将蛋白质消化成相对较小的片段，最后用 MALDI-TOF 质谱（框 8-2）确定这些蛋白质片段，以确定整条蛋白质的序列。

> Gavin 等（2002）的实验数据可以从 http://yeast.cellzome.com 上获得。Ho 等（2002）的实验数据可以从 http://mdsp.com/yeast 上获得。它们的数据都已经存储在了 BIND（生物分子相互作用网络数据库，biomolecular interaction network database）中了。想要获得更多蛋白质相互作用的数据库，可以参见“8.7 网络资源”中的表 8-18。关于假阳性和假阴性结果的问题，Schachter（2002）在论述酵母双杂交筛选的时候讨论过（见下文）。

运用这个策略，Gavin 等（2002）得到了 1167 个酵母种系表达的做了标记的蛋白，这些标记蛋白中有 589 个被纯化出来，最后总共鉴定出了 232 个蛋白质复合物。Ho 等（2002）挑选了 725 个“诱饵”蛋白并且鉴定出了数千个蛋白质-蛋白质结合物。在上述两个实验中鉴定出来的蛋白质复合物中包含了很多以前不知道功能的蛋白质，这更加突出了大规模高通量的方法的优势。

这种实验方法是建立在很多实验假设之上的，包括造成假阳性结果（生物学上不显著的相互作用）和假阴性结果（结果中丢失的生物学上有显著性的相互作用）的原因。导致出现假阴性结果的原因有如下几点。

① 将“诱饵”基因导入酵母细胞时，导入的位置一定要合适。不合适的位置造成“诱饵”蛋白不能在固有的细胞环境下发挥作用，是解释一些以前就已经知道的相互作用在实验中反而没有观察到的原因。

② 所加的亲和标记不能影响“诱饵”蛋白的功能。有时候，相对于蛋白质的平均大小（约 50ku），标记分子显得太大。

③ 有些瞬时的蛋白质相互作用可能被忽略掉了。

④ 有些蛋白质复合物要在高度特定的生理条件下才能形成，因此在实验中可能被忽略掉。

⑤ 疏水性蛋白质（这些蛋白质比可溶性蛋白质更难纯化）和小分子量蛋白质在实验中不容易被检测到。

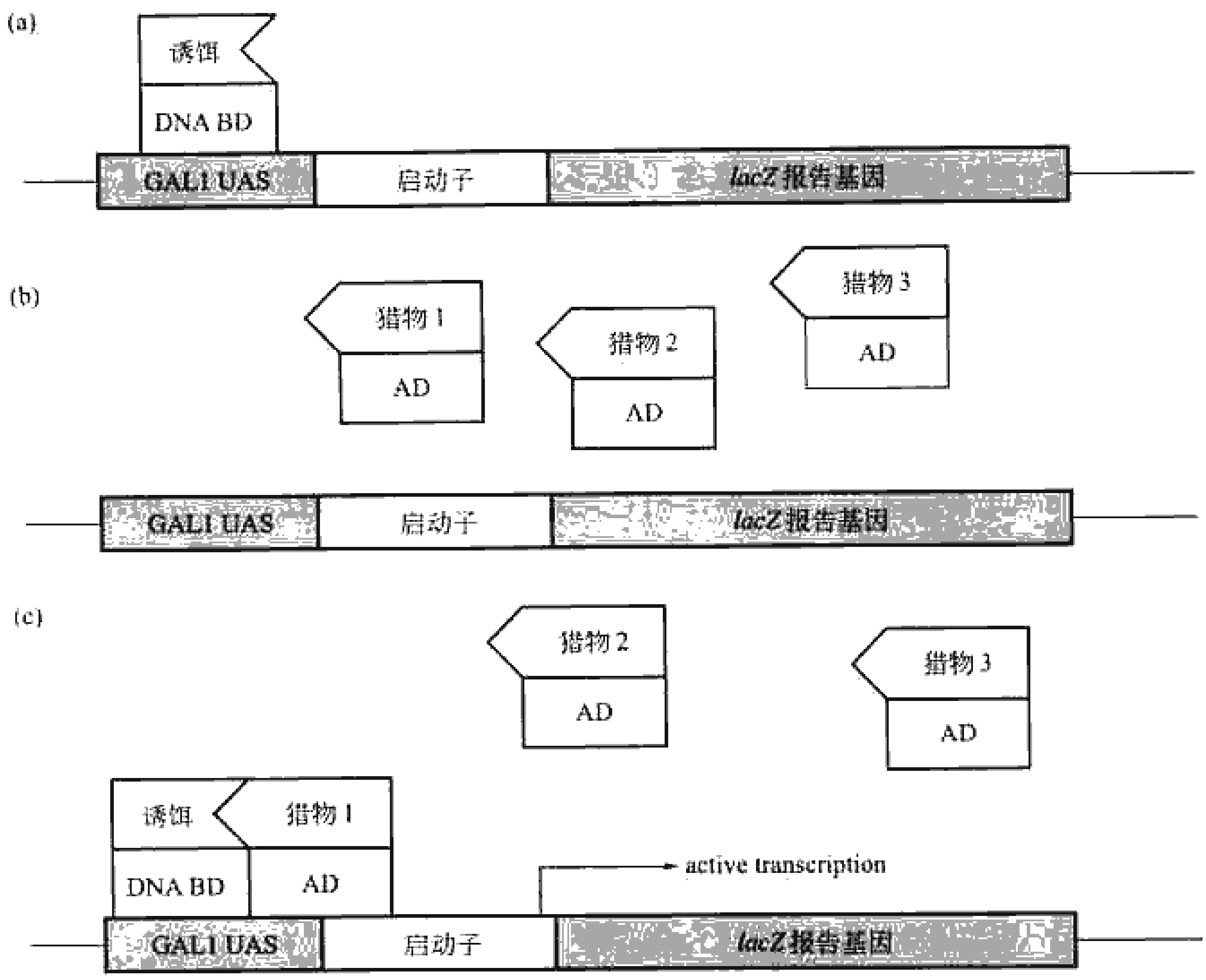
造成假阳性结果有很多可能的原因。有些蛋白本身固有的性质（如有些蛋白质具有“黏性”）使它们倾向于与其他蛋白质产生非特异的相互作用。变性的蛋白质也容易与其他蛋白产生非特异性的结合。

8.3.3 酵母双杂交系统

酵母双杂交是检测蛋白质-蛋白质相互作用的一种方法（Fields 和 Song，1989）。这种方法具有很强的通用性，已被用于检测很多物种中的蛋白质-蛋白质相互作用。酵母双杂交方法建立在如下事实上：酵母中的转录激活子 *GAL4* 由两个功能和结构上都相互独立的结构域组成（框 8-3）。cDNA 是编码一个我们感兴趣的蛋白（“诱饵”蛋白）和 *GAL4* 的 DNA

框 8-3 酵母双杂交系统

酵母双杂交系统可以用来测定蛋白质之间的相互作用。编码待研究蛋白质（如 huntingtin 蛋白质，Huntington 疾病会伴随这个蛋白的突变）的 cDNA 被用作“诱饵”来筛选编码了所有人的大脑中表达的蛋白质的 cDNA 文库，找到哪些蛋白质与待研究的蛋白质有相互作用。一段包含有 huntingtin cDNA 并且融合了 DNA 结合结构域（DNA binding domain，简称 BD）的 DNA 片段被引入酵母细胞中。BD 会与酵母基因组上的 *GAL4* 上游激活序列（UAS）结合，但是在没有合适的激活结构域（activator domain，简称 AD）时，乳糖操纵子基因将不会被激活（见下文）。含有数以千计的 cDNA 的文库被建立起来，每一个 cDNA 都与一个 AD 序列进行融合。当库中的一个克隆（融合了 AD 的“猎物 1”蛋白）与融合了 BD 的“诱饵”蛋白相互结合时，激活结构域将能够让下游的乳糖操纵子报告基因激活。酵母细胞中的 DNA 能够被鉴定出来，我们将编码“猎物 1”的 DNA 测序。对于一个特定蛋白可能会鉴定出多个蛋白与它有相互作用。这个技术的一项应用就是 Li 等（1995）鉴定出了 huntingtin 结合蛋白（huntingtin-associated protein，简称 HAP-1）。这种蛋白在大脑中大量存在被认为是造成 Hingtington 疾病中常见的多聚谷氨酰胺扩展现象的一个可能的原因（见第 18 章）。



结合结构域的融合体。然后一个 cDNA 集合（“猎物”蛋白构成的库）被克隆到含有 *GAL4* 激活结构域的载体中。单独一个 GAL4 DNA 结合结构域是不能激活转录的，但是当“诱饵”蛋白与另外一个由 cDNA 库中 cDNA 表达的蛋白相互作用（物理接触）时，这两个蛋白质的亲密结合使得 *GAL4* 的报告基因能够被转录。

除了用一个“诱饵”蛋白去筛选一个库的策略之外，酵母双杂交系统还经常被用来检测一个已知的“诱饵”蛋白和一个单独被克隆的“猎物”蛋白之间的相互作用。利用这种方法，很多对蛋白质-蛋白质相互作用都被成功地检测出来。与利用“诱饵”蛋白筛选 cDNA 文库的方法相比，这种只检测一个“猎物”蛋白的方法具有一个优势，

就是它可以很系统地检测一对有可能发生相互作用的蛋白质之间是否发生相互作用，它的劣势就是无法发现新的发生相互作用的蛋白质对，而利用"诱饵"蛋白筛选 cDNA 文库的方法就可能做到这一点。

芽殖酵母（*S. cerevisiae*）中，所有本质上有可能发生相互作用的蛋白质对都已经用酵母双杂交的方法分析过了。Uetz 等（2000）报告了实验中发现的 957 个相互作用，这 957 个相互作用中包含了 1004 个酵母蛋白。而 Ito 等（2001）则确定了 3278 个酵母蛋白之间的 4549 个相互作用。这些数据集对于我们发现可能存在的由蛋白质和它们之间的相互作用构成的通路（pathway）很有用处。但令人感到惊奇的是，这两个数据集只有 20%的重叠数据。这两个数据集之间的不协调很可能是由于实验是在不同的生理条件下完成的，或者是因为假阳性和假阴性的错误结果来源于不同的方面。

Ito 等（2001）的数据集可以从 http://genome.c.kanazawa-u.ac.jp/Y2H 上获得。此外，酵母双杂交技术在整个蛋白质组水平上运用的另一个例子就是 Rain 等（2001）为 Helicobacter pylori 绘制了一张很大的蛋白质-蛋白质相互作用图谱。

酵母双杂交策略与亲和层析方法最大的不同是酵母双杂交系统只能被用来检测两个蛋白质之间的相互作用。而相反地，亲和层析的方法则可以分离和检测由多个蛋白质构成的集团。

和其他的高通量筛选技术一样，酵母双杂交技术也存在很多与假阳性和假阴性结果有关的问题（如可信度和覆盖率）。除了前文在论述亲和层析方法时提到的问题外，对于酵母双杂交，如果蛋白质-蛋白质相互作用不是出现在一个专门的细胞核环境中时，可能会导致假阴性的结果。如果"诱饵"蛋白没有与"猎物"蛋白结合就自动转录了一个报告基因，则可能会出现假阳性的结果。一些黏性的"猎物"蛋白可能会非特异地结合到很多"诱饵"蛋白上，这也会导致假阳性的结果。对酵母双杂交实验结果的仔细分析可以使假阳性率和假阴性率降低，如确定相互混杂结合的蛋白质。

8.3.4 Rosetta Stone 方法

一些研究小组已经开始采用计算的方法大规模预测蛋白质的功能。Marcotte 和他的同事（1999a，1999b）以及 Enright 等（1999）提出了这样一个假说：一些编码在一个基因组不同基因中的蛋白对，在另外一个基因组中却融合（fuse）成了一个蛋白（图 8-23）。Marcotte 等分析了多个基因组，在大肠杆菌（*E. coli*）中鉴定了 6809 个这样的例子，在酵母（*S. cerevisiae*）中鉴定了 45502 个这样的例子。这种对蛋白质结构域的分析被称为 Rosetta

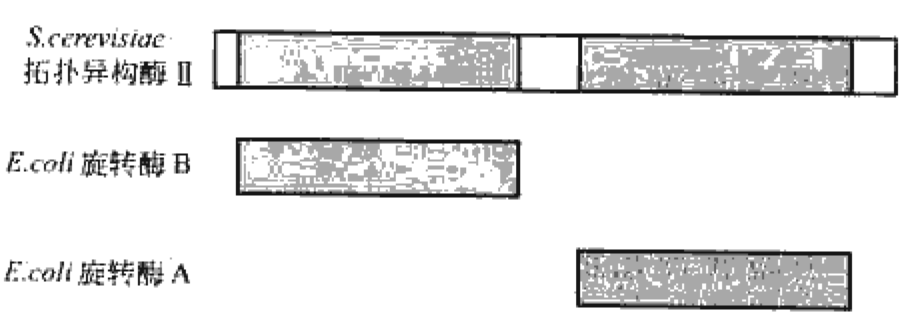


图 8-23 Rosetta Stone 方法可以以基因组序列信息为基础来预测蛋白质之间的功能联系。在一个物种的基因组序列中搜索单独存在的基因，如果在另外一个物种中这两个基因融合在一起则说明这两个蛋白质是一个功能相关的通路的一部分。因为基因融合后的物种由于共调控和两个蛋白可同时拥有较高的局部浓度而具有进化上的优势

Stone方法。Rosetta Stone方法假设两个发生融合的蛋白质具有相关的生物学功能，例如它们可能在同一个蛋白质复合物（complex）、通路（pathway）或过程（process）中发挥功能。这种方法还可以预测可能存在的蛋白质-蛋白质相互作用。因为融合后的蛋白可以保证在一个局部区域具有更为有效的蛋白浓度，所以发生了基因融合事件的物种可以因此而受益。当然也不能排除除了功能相关的原因外还有其他原因导致蛋白质融合事件的发生。

"Rosetta Stone"是指古代埃及的一种石碑，上面用象形文字、通俗的希腊文和经典的希腊文记录着一段铭文，且这三种文字记录的是同一篇文献，如果它们同时被发现象形文字就可能被破解。

8.4 分析细胞通路的生物信息学方法

细胞中众多蛋白质功能的信息可以被整合在数据库中，并且以蛋白质网络图（protein network map）的形式显示出来（Schächter，2002）。一个通路（pathway）就是一个相互联系在一起的生化反应的集合（Karp，2001）。将生化反应组织成通路的目的是为了更好地展现细胞中复杂的生物过程。

哪些数据可以用于通路作图（pathway maps）呢？从一些实验数据（如酵母双杂交）或是计算结果（如Rosetta Stone）中，我们可以得到整个蛋白质组水平上的蛋白质-蛋白质相互作用。除了整个蛋白质之间的相互作用以外，基于一些筛选实验（包括酵母双杂交系统）得到的蛋白质中特定结构域之间相互作用的信息也在不断地积累。

一些网页服务器提供了通路作图的服务（表8-18列出了这些网页资源）。Boehringer Mannheim公司创作了一种便于理解的显示生化反应通路的图表，这些图表可以在ExPASy上通过在线的方式获得（图8-24）。

EcoCyc/MetaCyc数据库提供了很多物种的通路图。EcoCyc是关于模式生物大肠杆菌（*E. coli*）K12品系的数据库（Karp等，2002a）（见第14章）。MetaCyc中包含了450个通路和1100个分布在158个物种中的酶（Karp等，2002b）。EcoCyc/MetaCyc有多种查询选项供选择（图8-25），并且对各个物种的数据都有总结（图8-26）。我们可以查看任意一条特定的通路，这时页面上会显示基因、它们编码的蛋白质以及回路中代谢物或其他参与分子的信息（图8-27～图8-29）。

蛋白质相互作用网络经常被外推到其他物种上。研究酵母时所得到的网络信息经常被用来推断其他物种中类似的网络，例如小鼠和人。利用这种方法的一个例子就是，Paley和Karp（2002）利用EcoCyc和MetaCyc数据库中有关大肠杆菌（*E. coli*）代谢通路（pathway）的信息来预测*H. pylori*中对应的代谢通路。他们发现通过算法预测的结果和那些经过专家人工分析预测出的结果有一部分是重叠的。每一种预测方法都有假阳性和假阴性。不同物种间的通路（pathway）比较的一个基本假设就是直系同源蛋白可以被可靠地确定。

另一个关于通路的数据库是kyoto encyclopedia of genes and genomes（KEGG）（Kanehisa和Goto，2002）。KEGG数据库包含不同物种之间同源蛋白的注释和数以百计的代谢（或其他）通路的信息。

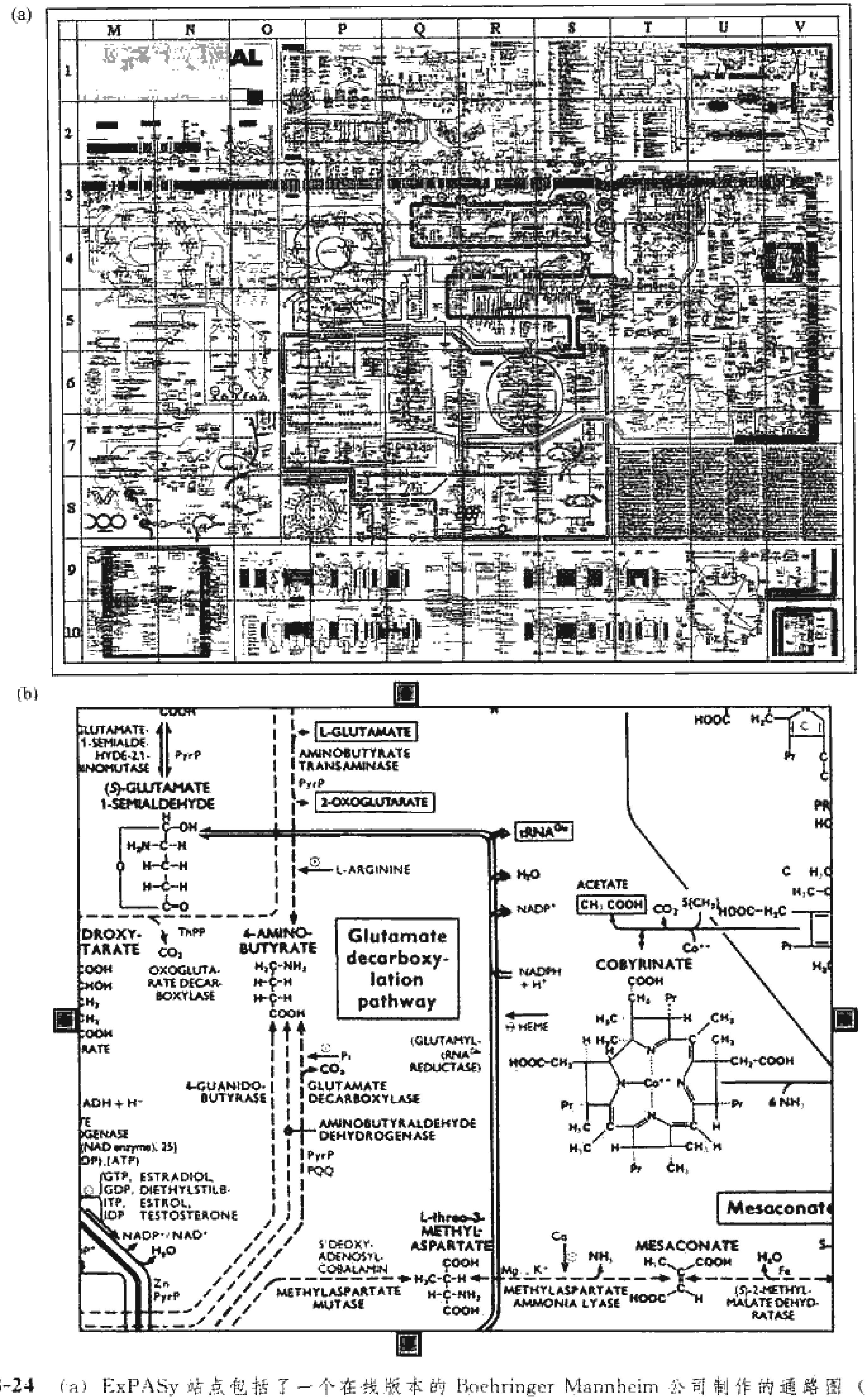


图 8-24　(a) ExPASy 站点包括了一个在线版本的 Boehringer Mannheim 公司制作的通路图 (pathway maps)。(b) 在站点中搜索谷氨酸（大脑中主要的兴奋型氨基酸神经递质）会得到一个图显示这个通路的详细情况。可以从 http://biochem.boehringer-mannheim.com/prodinfo.fst.htm?/techserv/metmap/.htm 上获得该图

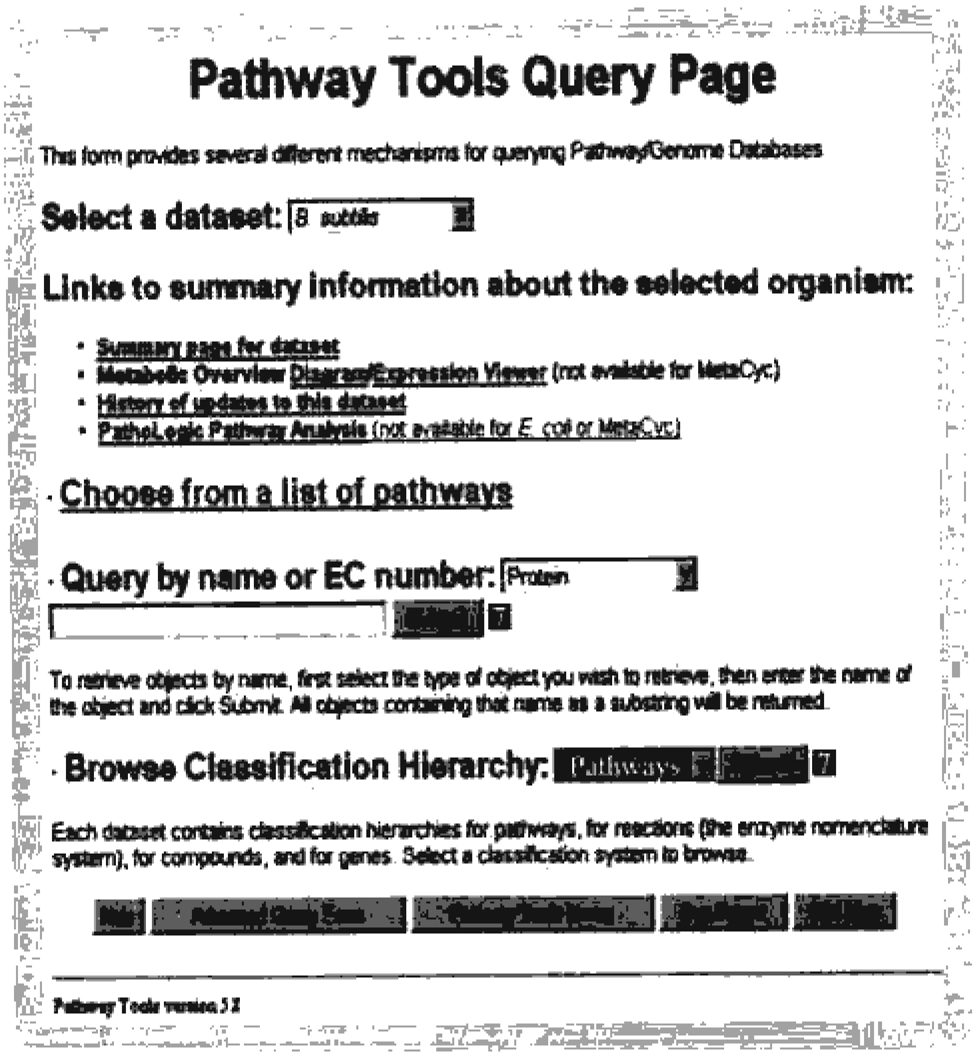


图 8-25 EcoCyc/MetaCyc搜索页面提供了访问pathways数据库的接口。这个数据库中收录了很多物种通路的信息（物种包括 *Bacillus subtilis*、*Chlamydiatrachomatis*、*E. coli*、*Helicobacter pulori*、*Haemophilus influenzae*、*Mycobacterium pneumoniae*、*Pseudomonas aeruginosa*、*S. cerevisiae*、*Treponema pallidum*）

Summary: The primary data source for this dataset was the full genome sequence of Saccharomyces cerevisiae [1]. This dataset was derived computationally using the PathoLogic program [2], with some additional manual curation. Manual curation has not occurred since 1998. Note that each YeastCyc gene contains a WWW link to the SGD database. Please refer to SGD for more detailed information about yeast gene products.

Genetic Element	Total Genes	Protein Genes	RNA Genes	Unidentified ORFs	Size (bp)
Chromosome I	111	107	4	54	230,209
Chromosome II	437	423	14	205	813,138
Chromosome III	225	214	11	130	315,341
Chromosome IV	826	796	30	409	1,531,974
Chromosome V	304	280	24	133	576,870
Chromosome VI	143	133	10	74	270,148
Chromosome VII	599	560	39	287	1,090,936
Chromosome VIII	290	279	11	141	562,638
Chromosome IX	233	223	10	122	439,885
Chromosome X	414	386	28	201	745,440
Chromosome XI	350	333	17	161	666,448
Chromosome XII	571	536	35	306	1,078,172
Chromosome XIII	507	486	21	256	924,430
Chromosome XIV	429	415	14	222	784,330
Chromosome XV	582	562	20	297	1,091,283
Chromosome XVI	505	487	18	267	948,061
Mitochondrial Chromosome	0	0	0	0	78,520
Total	6526	6220	306	3265	12,147,823

Pathways	87
Enzymatic Reactions	600
Transport Reactions	1
Polypeptides	6241
Protein Complexes	42
Enzymes	509
Transporters	3
Compounds	457
Transcription Units	0
tRNAs	275

图 8-26 EcoCyc数据库包括了对各个物种的总结信息。图中显示了芽殖酵母的一部分结果

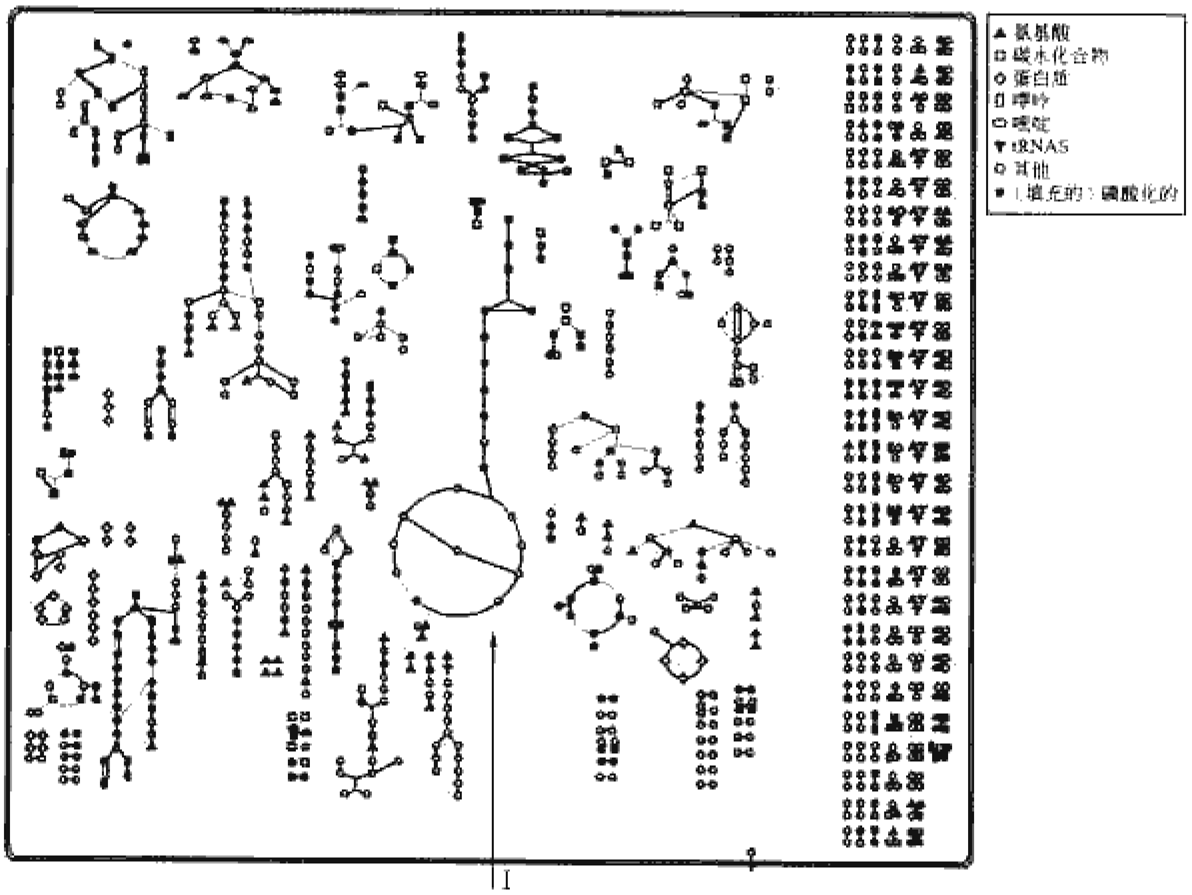


图 8-27 EcoCyc 数据库可以将整个代谢通路以图的形式显示出来。这里显示的是芽殖酵母的代谢通路。箭头所指的是三羧酸循环（箭头 I）

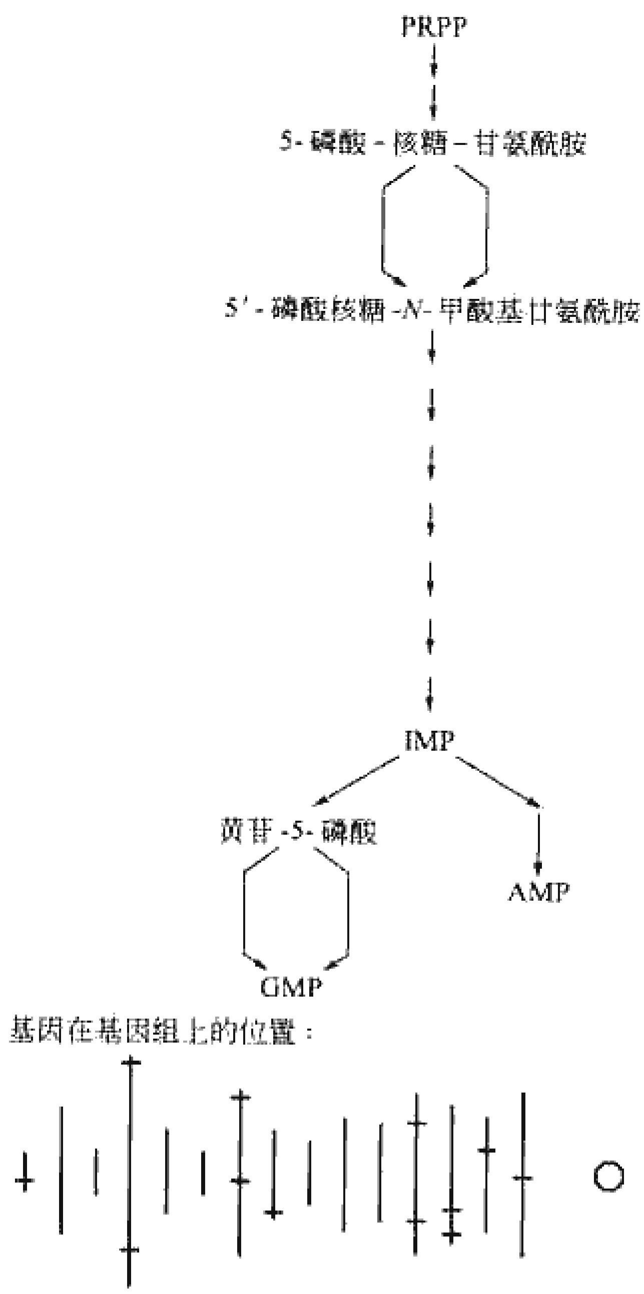


图 8-28 EcoCyc 数据库中的磷酸核糖焦磷酸通路

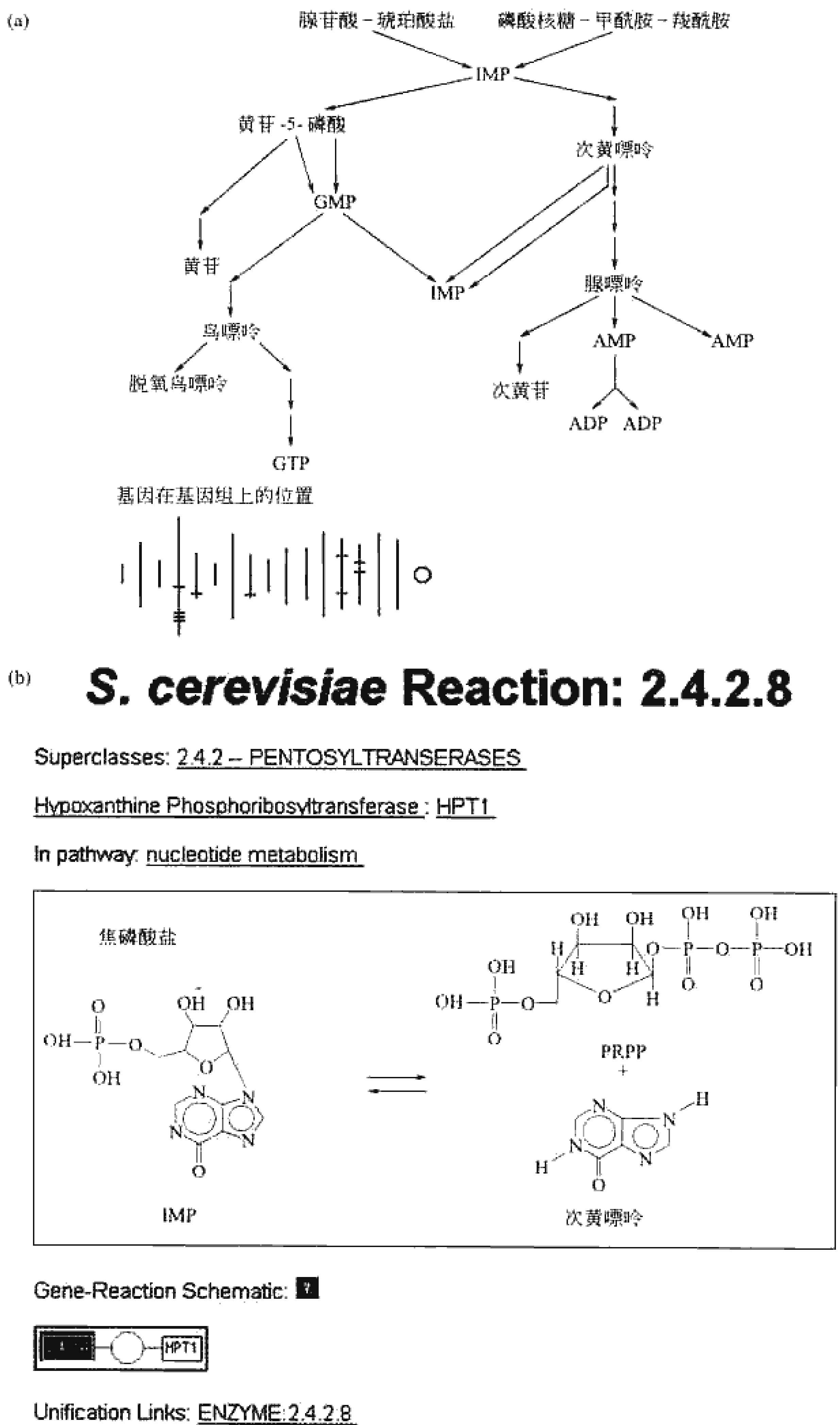


图 8-29 (a) EcoCyc 数据库收录了单个的通路。(b) 点击通路中的组分，将显示有关其底物和产物的详细信息

KEGG 数据库中有数百个代谢和调节通路图，表 8-10 列出了主要的分类。选择其中的一个分类，例如选择 "human neurodegenerative disorders"，我们可以看到对这条通路的描述：肌萎缩性侧部硬化（ALS；Lou Gehrig's disease）[图 8-30(a)]。超氧化物歧化酶基因 *SOD1* 的突变是引起这种疾病的一个常见的原因。*SOD1* 编码一种酶，这种酶在正常条件下可以将有毒的氧代谢产物超氧化物（O_2^-）转化为过氧化氢和水。如 KEGG 数据库中的通路图所显示的，现已发现 *SOD1* 与很多其他蛋白都有直接或间接的联系，这些蛋白中包括那些参与了细胞凋亡过程（apoptosis，或者叫细胞程序性死亡，programmed cell death）的蛋白。点击 SOD1 会进入一个页面，这个页面中有关于这个蛋白质以及其氨基酸序列的信息，还包括一些外部链接，如酶学委员会的编号等 [图 8-30(b)]。进一步点击链接会出现另外一些信息，包括用蛋白质预测工具预测出的跨膜结构域、蛋白质分类标记和模体 [图 8-30(c)]。

表 8-10　KEGG 数据库中代谢和调节通路的主要分类

这些分类中包括数百个代谢通路图	
代谢通路：	调节通路：遗传信息处理
碳水化合物代谢	分选和降解
能量代谢	复制和修复
脂类代谢	调节通路：环境信息处理
核苷酸代谢	膜转运
主要氨基酸代谢	信号转导
其他种类的氨基酸代谢	配体和受体相互作用
碳水化合物类复合物代谢	调节通路：细胞过程
脂类复合物代谢	细胞运动
辅因子和维生素代谢	细胞生长和死亡
二级代谢物的合成	细胞通讯
宾主共栖生物的降解	发育
调节通路：遗传信息处理	行为
转录	调节通路：人类疾病
翻译	神经退化性紊乱

注：资料来源于 http://www.genome.ad.jp/kegg/。

上面论述的 *SOD1* 的例子突出显示了 KEGG 数据库的价值。KEGG 数据库收录了大量的蛋白质过程和细胞过程，它的覆盖程度很广泛。从上面 *SOD1* 的例子还可以看出 KEGG 中描述的很多细胞过程都是物种特异的。有一点需要注意，KEGG 数据库主要是基于细菌基因组和通路数据建立的，因此 KEGG 中的数据不能简单地运用到真核生物中去，我们在使用 KEGG 数据库的时候要小心。

可以通过 http://ecocyc.org 来访问 EcoCyc 通路数据库。MetaCyc 的网址是 http://ecocyc.org/ecocyc/metacyc.html。

H. pylori **通路数据库的网址是 http://ecocyc.org。**

KEGG 数据库的网址是 http://www.genome.ad.jp/kegg/。第 23 个发行版本（2002 年 7 月）中包括约 350000 个基因，这些基因来自 14 种真核生物（包括人、小鼠、大鼠）和 88 种细菌和古细菌。

从微阵列实验中获得的基因表达数据可以标记上 KEGG 中通路的编号，然后将表达值显示在 KEGG 的通路图中。DRAGON 数据库提供了这个功能（http://pevsnerlab.kennedykrieger.org）。

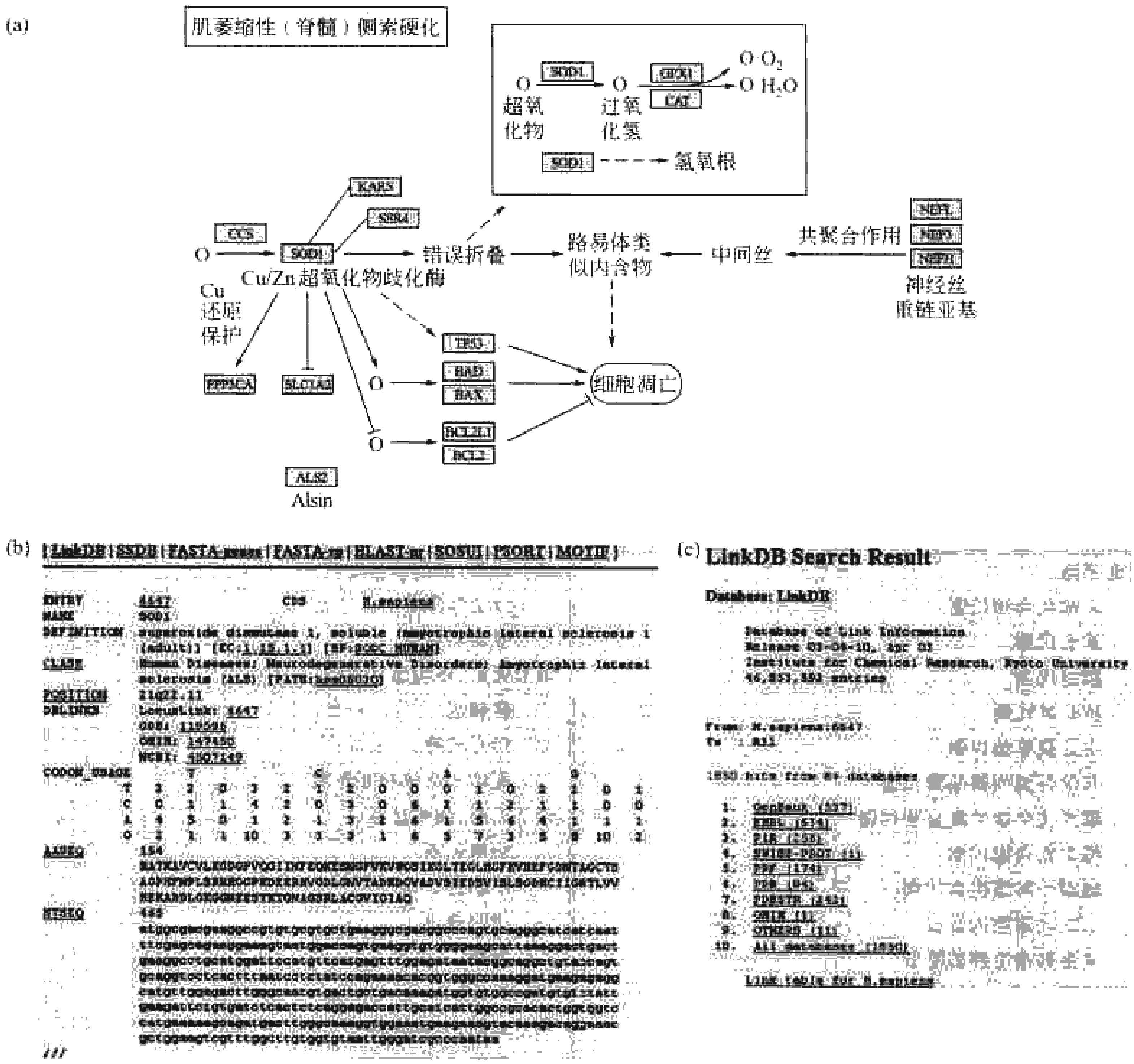


图 8-30 KEGG数据库收录了与100多个完整基因组中的350000个基因和蛋白质有关的通路图，其中有一个分类是针对疾病的。图（a）中显示的是与肌萎缩性侧部硬化（ALS；Lou Gehrig's disease）有关的通路。超氧化物歧化酶基因 *SOD1* 与很多家族性 ALS 病例有关；（b）点击它的小图标将进入 KEGG 数据库中 *SOD1* 的登录页面，（c）并且给出了本章介绍的各种蛋白质预测软件的链接

8.5 展望

在过去的几十年里，我们对蛋白质性质的理解已经发生了突飞猛进的变化，从生化功能这个层次到蛋白质在细胞过程中所起的作用皆是如此。现在人们可以轻易地获取很多基于网页的用来评价单个蛋白质生化性质的工具。此外，一些高通量的方法（如二维凝胶电泳和酵母双杂交）已经被用来定义蛋白质组内所有蛋白质的功能。现在，还有很多未知功能的蛋白，因为它们没有已经被确定了功能的同源序列。将各种高通量实验策略运用在模式生物上，我们还将继续获得更多更广泛的有关蛋白质功能的描述，例如大规模的蛋白质定位信息和蛋白质相互作用信息。

8.6 常见问题

很多用于研究蛋白质功能的实验或计算的策略都存在局限。二维凝胶电泳可以很好地研究那些表达量相对较高的蛋白质的性质，但是数以千计的蛋白质在正常细胞中的表达水平都非常低，二维凝胶电泳对确定这些蛋白质的性质显然无能为力。酵母双杂交的方法存在太多假阴性的结果。如果要检测的蛋白质不能在酵母细胞中表达或者因为重组融合的办法导致要研究的蛋白质性质发生了变化，都会造成假阴性的结果，致使无法检测两个蛋白质之间的相互作用。

用高通量的手段给蛋白质功能所做的划分只能被视为一个试验性的假设。还需要常规的细胞和分子生物学实验来确定这些蛋白质的功能。

8.7 网络资源

网络资源参见表 8-11～表 8-18。

表 8-11 分析蛋白质模体的工具

程 序	注 释	网 址
InterProScan	在 EBI	http://www.ebi.ac.uk/interpro/scan.html
Ppsearch	在 EBI	http://www2.ebi.ac.uk/ppsearch/
PRATT	在 EBI	http://www2.ebi.ac.uk/pratt/
ProfileScan Server	在 ISREC	http://hits.isb-sib.ch/cgi-bin/hits.motifscan
PROSCAN(PROSITE SCAN)	在 PBIL	http://npsa-pbil.ibcp.fr/cgi-bin/npsa.automat.pl?page=npsa.prosite.html
ScanProsite tool	在 ExPASy	http://ca.expasy.org/tools/scnpsite.html
SMART	在 EMBL	http://smart.embl-heidelberg.de/
TEIRESIAS	在 IBM	http://cbcsrv.watson.ibm.com/Tspd.html

注：资料来自 ExPASy（http://www.expasy.org/tools/）。

表 8-12 分析蛋白质一级结构或二级结构性质的工具

程 序	来源/注释	网 址
COILS	预测蛋白质中的螺旋卷曲区域	http://www.ch.embnet.org/software/COILS_form.html
Compute pI/Mw	由 ExPASy 提供	http://www.expasy.org/tools/pi_tool.html
Drawhca	疏水性聚类分析	http://psb00.snv.jussieu.fr/hca-seq.html
Helical Wheel	绘制 helical wheel，如对于一个给定的序列，绘制 α 螺旋在一个轴的投影	http://www.site.uottawa.ca/~turcotte/resources/HelixWheel
M.M., pI, composition titrage	由 Aterlier Bio Informatique de Marseille 提供	http://www.up.univ-mrs.fr/~wabim/d_abim/compo-p.html
Paircoil	预测蛋白质中的螺旋卷曲区域	http://nightingale.lcs.mit.edu/cgi-bin/score
PeptideMass	由 ExPASy 提供	http://kr.expasy.org/tools/peptide-mass.html
REP	寻找一个蛋白质序列中的重复区	http://www.embl-heidelberg.de/~andrade/papers/rep/search.html
SAPS	对蛋白质序列进行统计分析	http://www.isrec.isb-sib.ch/software/SAPS_form.html

注：资料来自 ExPASy（http://www.expasy.org/tools/）。

表 8-13 确定蛋白质上糖基化位点的网页资源

程序	注释/来源	网址
DictyOGlyc 1.1 prediction Server	用神经网络算法预测盘基网柄菌(*Dictyostelium discoideum*)中蛋白质上的O型N-乙酰葡萄糖胺糖基化位点	http://www.cbs.dtu.dk/services.DictyOGlyc
NetOGlyc	预测哺乳动物蛋白上的O型糖基化位点	http://www.cbs.dtu.dk/services/NetOGlyc/
YinOYang 1.2	用神经网络预测真核生物蛋白序列上的O型N-乙酰葡萄糖胺糖基化位点	http://www.cbs.dtu.dk/services/YinOYang/

表 8-14 分析蛋白质翻译后修饰的工具

程序	注释	网址
Big-PI Predictor	预测GPI修饰位点	http://mendel.imp.univie.ac.at/gpi/
DGPI	发现/预测蛋白质上的GPI断裂位点(GPI-anchor)	http://129.194.186.123/GPI-anchor/index.en.html
NetPhos 2.0 Prediction Server	用神经网络方法预测真核生物蛋白质中丝氨酸(serine)、苏氨酸(threonine)、酪氨酸(tyrosine)残基上的磷酸化位点	http://www.cbs.dtu.dk/services/netPhos/
Sulfinator	预测酪氨酸(tyrosine)残基上的硫化位点	http://www.expasy.org/tools/sulfinator

表 8-15 特定氨基酸出现频率高的蛋白质

氨基酸	蛋白质
C	富含二硫键的蛋白质，金属硫蛋白，锌指蛋白
D,E	酸性蛋白质
G	胶原蛋白
H	组蛋白，富含组氨酸的糖蛋白
W,L,P,Y,I,V,M,A	跨膜蛋白质
K,R	核蛋白(核定位信号)
N	网柄菌属中的蛋白质
P	胶原蛋白，纤维蛋白，SH3/WW/EVHI结合位点
Q	三联体密码错乱突变基因编码的蛋白质
S,R	一些RNA结合模体
S,T	黏蛋白，寡聚糖结合位点
abcdefg	七个氨基酸残基为一个周期螺旋(a和d是疏水氨基酸残基)，如肌凝蛋白

表 8-16 基于网页的预测蛋白质细胞定位的程序

程序	注释	网址
ChloroP	预测蛋白质序列中是否有叶绿体转运多肽(chloroplast transit peptides，简称cTP)存在的软件	http://www.cbs.dtu.dk/services/ChloroP/
MITOPROT	计算蛋白质的氨基末端是否有线粒体定位序列，以及该定位序列的断裂位点的软件	http://www.mips.biochem.mpg.de/cgi-bin/proj/medgen/mitofilter/
PSORT	预测蛋白质的分选信号及细胞定位的软件	http://psort.ibb.ac.jp/
SignalP	预测原核生物和真核生物信号肽是否存在以及信号肽断裂位点的软件	http://www.cbs.dtu.dk/services/SignalP/
TargetP	预测真核生物蛋白质序列亚细胞定位的软件	http://www.cbs.dtu.dk/services/TargetP/

表 8-17 预测蛋白质跨膜区结构域的服务器

程序	注释/来源	网址
DAS Server	—	http://www.sbc.su.se/~miklos/DAS/
HMMTOP	预测跨膜螺旋和蛋白质的拓扑结构	http://www.enzim.hu/hmmmtop/
PredictProtein server	预测跨膜螺旋的位置和拓扑结构	http://dodo.cpmc.columbia.edu/predictprotein/
SOSUI.	跨膜蛋白的分类和二级结构预测	http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html
TMpred	—	http://www.ch.embnet.org/software/TMPRED_form.html
TMHMM(v.2.0)	由丹麦科技大学生物序列分析中心维护	http://www.cbs.dtu.dk/services/TMHMM-2.0/
TopPred2	蛋白质的拓扑结构预测	http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html

注：来源于 ExPASy 网站。

表 8-18 基于网页的蛋白质-蛋白质相互作用数据库和蛋白质-配体相互作用数据库

数据库	注释	网址
BIND(生物分子相互作用网络数据库，The Biomolecular Interaction Network Database)	存储了相互作用，分子复合物和通路信息的数据库	http://www.bind.ca.
Cellzome	—	http://yeast.cellzome.com
DIP(相互作用蛋白数据库，The Database of Interacting Proteins)		http://dip.doe-mbi.ucla.edu/main.html
DLRP(配体受体数据库，Database of Ligand-Receptor Partners)	一个收录有已知的蛋白质受体和配体对的数据库	http://dip-doe-mbi.ucla.edu/main.html
FlyBase	见 Jacq(2001)	http://fly.ebi.ac.uk:7081
FlyNets	见 Jacq(2001)	http://gifts.univ-mrs.fr/FlyNets
KEGG(京都基因与基因组百科全书，Kyoto Encyclopedia of Genes and Genomes)	—	http://www.genome.ad.jp/kegg
GeNet(基因网络数据库，Gene Networks database)	见 Jacq(2001)	http://www.csa.ru: 85/inst/gorb.dep/inbios/genet/genet.htm
STKE(信号转导知识数据库，Signal Transduction Knowledge Environment)	见 Jacq(2001)	http://www.stke.org
Transfac(转录因子数据库，Transcription Factor database)	—	http://transfac.gbf-braunschweig.de/TRANSFAC/index.html
YPD，PomberPD，WormPD	蛋白质组数据库	http://www.proteome.com/databases/index.html

注：来源于 http://dip.doe-mbi.ucla.edu/databases.html，有些数据库在 Schächter (2002) 的文章上有列表。

8.8 问题讨论

1. InterPro 数据库是一个很重要的资源，它整理了各个不同的数据库关于蛋白质签名区（signature）的信息。如果有多个数据库对同一个蛋白质或同一个签名区都有描述，能否从 InterPro 上获取各种不同信息？InterPro 上的这些信息是否冗余？

2. 对于诸如酵母双杂交这样的高通量筛选技术有哪些主要的假设？概括地说，一般一个蛋白质有几个结合对象，在怎样的生理条件下发生结合？

8.9 问题

1. 大马哈鱼身体的颜色略带粉红色，而一些龙虾的颜色则是蓝色的（火烤时就会变成

红色），这是因为龙虾体内有一种被称为虾青素的着色团，它会结合到一种叫做虾青蛋白的载体蛋白上。请从 European lobster Homarus gammarus 上检索出虾青蛋白的序列，并指出这个蛋白有哪些物理性质（如分子量、等电点等）。利用 ExPASy 网站上的工具检测虾青蛋白是否包含有一些功能已知的结构域或模体能够解释它为什么能够与着色团结合？（要想获取更多关于虾青蛋白的性质，可以参考 ExPASy 上的一篇文章：http://www.expasy.org/spotlight/articles/sptlt026.html.）

2. 利用 ExPASy 站点上的工具评价一下人类突触融合蛋白（syntaxin）的性质。它是否含有螺旋卷曲区域？它有几个被预测出的跨膜结构域？它的功能是什么？先利用 ExPASy 的序列提取系统。

3. 嗅觉受体与视紫红质类似（rhodopsin-like）的 G 蛋白偶联受体（G-protein coupled receptor，简称 GPCR）蛋白超家族有关。利用 EBI 上的蛋白质组工具（http://www.ebi.ac.uk/proteome/），检测一下与这个蛋白质超家族有关的蛋白占小鼠的蛋白组的百分比，它们又占人类蛋白质组的百分比。

4. 利用 http://www.ebi.ac.uk/proteome/上的蛋白质组工具检测 *E. coli* K12 中最常见的 15 个蛋白质结构域是否也出现在人类的蛋白质组中。

8.10 自测题

1. 同一个结构域可能出现在一个蛋白质的氨基末端，也可能出现在另一个蛋白质的羧基末端，是吗？

(a) 是；

(b) 不是

2. 一般说来，蛋白质结构域和模式（也称为模体或指纹）的大小的关系是

(a) 它们一样长；

(b) 模式比结构域长；

(c) 结构域比模式长；

(d) 只能对于特定的蛋白质才能进行比较

3. 氨基酸序列［ST］X［RK］是蛋白激酶 C 底物磷酸化的保守位点附近的序列，这段氨基酸序列是

(a) 一个模体，可以确定一组同源蛋白；

(b) 一个模体，但不足以确定一组同源蛋白；

(c) 一个结构域，可以确定一组同源蛋白；

(d) 一个结构域，但不足以确定一组同源蛋白

4. 如果用软件预测一个尚不知道功能的蛋白的糖基化位点、硫化位点、磷酸化位点是其他翻译后修饰位点：

(a) 软件预测的结果多半不可靠；

(b) 因为不知道实际情况，所以很难评价；

(c) 软件能够预测相关修饰是否存在，但无法预测它们的生物学功能，只有通过实验确定；

(d) 软件能够预测相关修饰是否存在，通过实验确定它们的生物学功能的方法也不可行

5. Gene Ontology 的定义有一个潜在的假设，即根据分子功能、生物过程和细胞成分三个方面对基因或基因产物的描述：

(a) 对不同的物种都是相同的，不管是植物、蠕虫还是人类；

(b) 不同物种之间的差异很大；

(c) 不同物种之间的差异可能很大也可能没有差异，对具体的每一个基因都要单独评价；

(d) 不同物种之间的差异可能很大也可

能没有差异，需要由专家来具体分析每一个基因

6. 蛋白质细胞定位的信息主要是在哪个 Gene Ontology 分类中进行描述的？

(a) 分子功能；

(b) 细胞组分；

(c) 细胞定位；

(d) 生物过程

7. 下面选项中，哪些是预测蛋白质功能的方法？

(a) 找到结构同源蛋白；

(b) 研究诱饵-猎物相互作用 (bait-prey interactions)；

(c) 确定蛋白质等电点；

(d) 上面的选项都是

8. 作为高通量蛋白质分析的一种手段，二维凝胶电泳的主要优势在于：

(a) 样品准备和二维凝胶电泳的过程是连续的，可以实现自动化；

(b) 二维凝胶电泳的结果中包含有数以千计的关于蛋白质等电点和分子大小的信息；

(c) 这项技术非常适合于检测那些含量特别低的蛋白质；

(d) 这项技术非常适合于检测疏水蛋白质

9. 造成诸如酵母双杂交和亲和层析等高通量的筛选技术假阳性结果的原因是

(a) 有些蛋白质本身具有黏性；

(b) 一些诱饵蛋白在被引入到细胞中时，定位发生了错误；

(c) 一些蛋白质复合物的存在是瞬时的；

(d) 亲和标签或抗原表位标签可能对蛋白质的相互作用有影响

10. 下面对细胞通路 (cellular pathway) 的表述中，最确切的一个是

(a) 生化通路 (biochemical pathways) 的方向一般都不确定；

(b) 通路图中不使用 Gene Ontology 定义的术语；

(c) 绘制通路图通常需要准确地确定直系同源蛋白，但是准确地确定直系同源蛋白有一定的困难；

(d) 通路图通常是根据原核生物的一些信息绘制的，关于真核生物的通路图还很有限

8.11 推荐读物

有很多描述蛋白质家族的优秀的文章，引起广泛兴趣的是 Steven Henikoff、Peer Bork 和他们的同事的文章 (1997)。还有 Russell Doolittle 和 Peer Bork 发表在 Scientific American 上的综述 (1993)。

Bernard Jarq 发表了一篇关于蛋白质功能的极其精彩的综述 (2001)。这篇综述讨论了蛋白质功能的复杂性和研究蛋白质功能的生物信息学工具。Jacp 建议从六个层次上来考虑一个蛋白质的功能，从蛋白质的结构到它在一个生物群体中所起的作用。

参考文献

Apweiler, R., et al. InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16**, 1145-1150 (2000).

Ashburner, M., et al. Creating the gene ontology resource: Design and implementation. *Genome Res.* **11**, 1425-1433 (2001).

Ashburner, M., et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-29 (2000).

Austen, B. M., and Westwood, O. M. *Protein Targeting and Secretion.* IRL Press, Oxford, 1991.

Bork, P., and Gibson, T. J. Applying motif and profile searches. *Methods Enzymol.* **266**, 162-184 (1996).

Bork, P., and Koonin, E. V. Protein sequence motifs. *Curr. Opin. Struct. Biol.* **6**, 366-376 (1996).

Cooper, T. G. *The Tools of Biochemistry.* Wiley, New York, 1977.

Copley, R. R., Schultz, J., Ponting, C. P., and Bork, P. Protein families in multicellular organisms. *Curr. Opin. Struct. Biol.* **9**, 408–415 (1999).

Doolittle, R. F. The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64**, 287–314 (1995).

Doolittle, R. F., and Bork, P. Evolutionarily mobile modules in proteins. *Sci. Am.* **269**, 50–56 (1993).

Dunn, M. J. (ed.) From Genome to Proteome: Advances in the Practice and Application of Proteomics. Wiley-VCH, New York 2000.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).

Farmer, T. B., and Caprioli, R. M. Determination of protein-protein interactions by matrix-assisted laser desorption/ionization mass spectrometry. *J. Mass Spectrom.* **33**, 697–704 (1998).

Fields, S., and Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245–246 (1989).

Fountoulakis, M., Schuller, E., Hardmeier, R., Berndt, P., and Lubec, G. Rat brain proteins: Two-dimensional protein database and variations in the expression level. *Electrophoresis* **20**, 3572–3579 (1999).

Frankel, A. D., and Young, J. A. HIV-1: Fifteen proteins and an RNA. *Annu. Rev. Biochem.* **67**, 1–25 (1998).

Gavin, A. C., et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).

Geppert, M., Goda, Y., Stevens, C. F., and Sudhof, T. C. The small GTP-binding protein Rab3A regulates a late step in synaptic vesicle fusion. *Nature* **387**, 810–814 (1997).

Grünenfelder, B., et al. Proteomic analysis of the bacterial cell cycle. *Proc. Natl. Acad. Sci. USA* **98**, 4681–4686 (2001).

Henikoff, S., et al. Gene families: The taxonomy of protein paralogs and chimeras. *Science* **278**, 609–614 (1997).

Ho, Y., et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).

Hoogland, C., et al. The SWISS-2DPAGE database: What has changed during the last year. *Nucleic Acids Res.* **27**, 289–291 (1999).

Ito, T., et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).

Jacq, B. Protein function from the perspective of molecular interactions and genetic networks. *Brief. Bioinform.* **2**, 38–50 (2001).

Jones, D.T. Protein structure prediction in genomics. *Brief. Bioinform.* **2**, 111–125 (2001).

Kallioniemi, O. P., Wagner, U., Kononen, J., Sauter, G. Tissue microarray technology for high-throughput molecular profiling of cancer. *Hum. Mol. Genet.* **10**, 657–662 (2001).

Kanehisa, M., and Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

Karp, P. D. Pathway databases: A case study in computational symbolic theories. *Science* **293**, 2040–2044 (2001).

Karp, P. D., et al. The EcoCyc Database. *Nucleic Acids Res.* **30**, 56–58 (2002a).

Karp, P. D., Riley, M., Paley, S. M., and Pellegrini-Toole, A. The MetaCyc Database. *Nucleic Acids Res.* **30**, 59–61 (2002b).

Kononen, J., Bubendorf, L., Kallioniemi, A., Barlund, M., Schraml, P., Leighton, S., Torhorst, J., Mihatsch, M. J., Sauter, G., Kallioniemi, O. P. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat. Med.* **4**, 844–847 (1998).

Kumar, A., and Snyder, M. Protein complexes take the bait. *Nature* **415**, 123–124 (2002).

Kumar, A., et al. Subcellular localization of the yeast proteome. *Genes Dev.* **16**, 707–719 (2002).

Langen, H., et al. Two-dimensional map of human brain proteins. *Electrophoresis* **20**, 907–916 (1999).

Li, X. J., et al. A huntingtin-associated protein enriched in brain with implications for pathology. *Nature* **378**, 398–402 (1995).

Lupas, A. Predicting coiled-coil regions in proteins. *Curr. Opin. Struct. Biol.* **7**, 388–393 (1997).

Lupas, A., Van Dyke, M., and Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164 (1991).

MacBeath, G. Protein microarrays and proteomics. *Nat. Genet.* Supplement **32**, 526–532 (2002).

Mann, G. *The Chemistry of the Proteids.* The Macmillan Company, New York, 1906.

Marcotte, E. M., et al. Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753 (1999a).

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., and Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999b).

Molloy, M. P., and Witzmann, F. A. Proteomics: Technologies and applications. *Briefings in Functional Genomics and Proteomics* **1**, 23–39 (2002).

Mulder, N. J., et al. InterPro: An integrated documentation resource for protein families, domains, and functional sites. *Brief. Bioinform.* **3**, 225–235 (2002).

Mulder, N. J., et al. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31**, 315–318 (2003).

O'Farrell, P. H. High resolution two-dimensional electrophoresis of proteins. *J Biol. Chem.* **250**, 4007–4021 (1975).

Østergaard, M., et al. Proteome profiling of bladder squamous cell carcinomas: Identification of markers that define their degree of differentiation. *Cancer Res.* **57**, 4111–4117 (1997).

Paley, S. M., and Karp, P. D. Evaluation of computational metabolic-pathway predictions for *Helicobacter pylori*. *Bioinformatics* **18**, 715–724 (2002).

Pevsner, J., Hou, V., Snowman, A. M., and Snyder, S. H. Odorant-binding protein. Characterization of ligand binding. *J. Biol. Chem.* **265**, 6118–6125 (1990).

Ponting, C. P. Issues in predicting protein function from sequence. *Brief. Bioinform.* **2**, 19–29 (2001).

Rain, J. C., et al. The protein–protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211–215 (2001).

Ratnam, M., Nguyen, D. L., Rivier, J., Sargent, P. B., and Lindstrom, J. Transmembrane topography of nicotinic acetylcholine receptor: Immunochemical tests contradict theoretical predictions based on hydrophobicity profiles. *Biochemistry* **25**, 2633–2643 (1986).

Sanchez, J. C., et al. The mouse SWISS-2D PAGE database: A tool for proteomics study of diabetes and obesity. *Proteomics* **1**, 136–163 (2001).

Schächter, V. Bioinformatics of Large-Scale Protein Interaction

Networks. *Computational Proteomics. A Supplement to BioTechniques*, 16–27 (2002).

Sigrist, C. J., et al. PROSITE: A documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* **3**, 265–274 (2002).

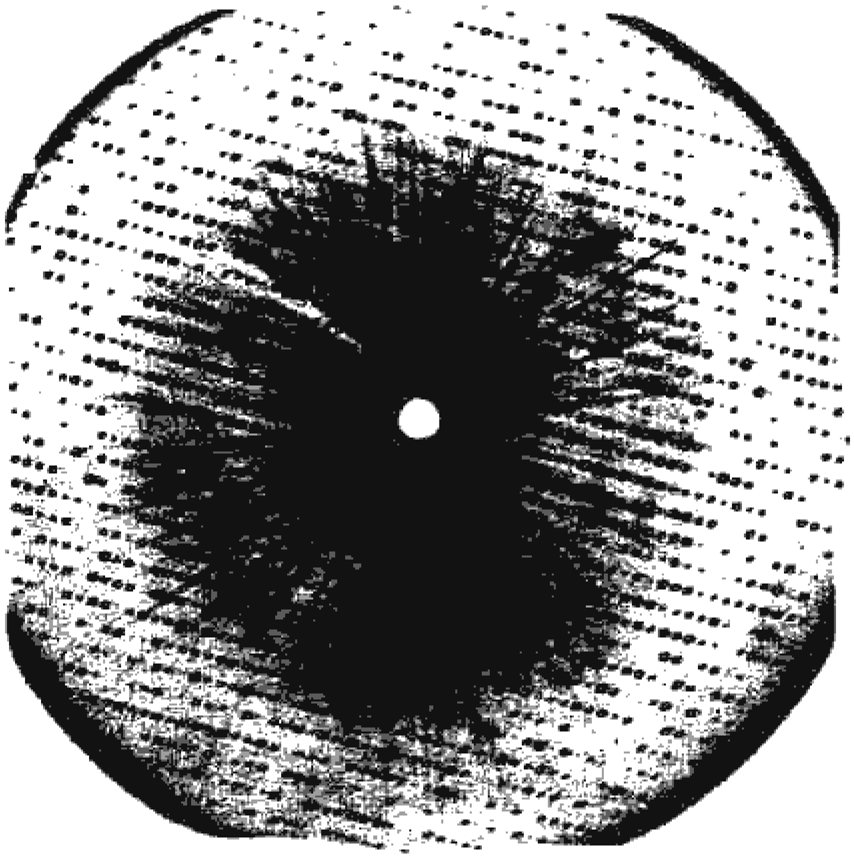
Sonnhammer, E. L., and Kahn, D. Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.* **3**, 482–492 (1994).

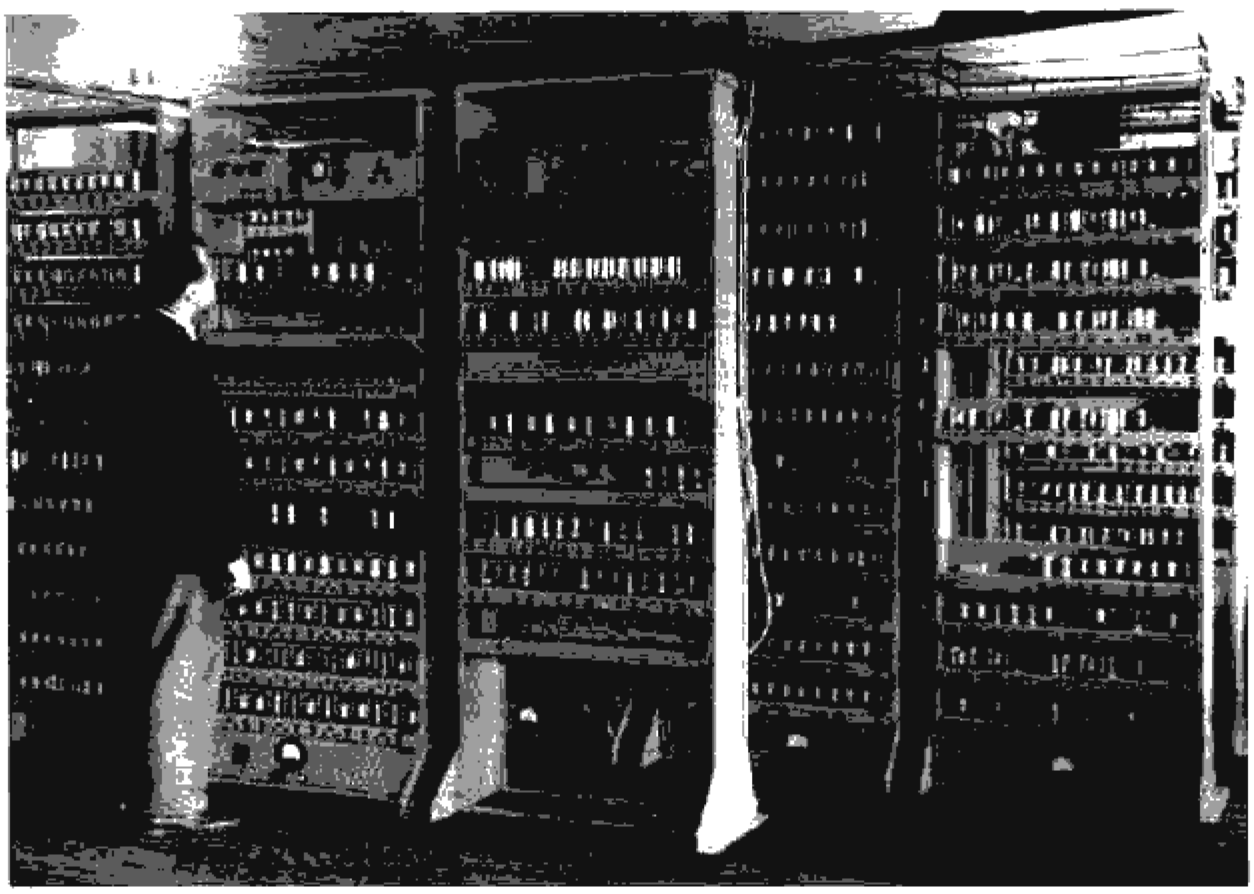
Stroud, R. M., and Walter, P. Signal sequence recognition and protein targeting. *Curr. Opin. Struct. Biol.* **9**, 754–759 (1999).

Takai, Y., Sasaki, T., and Matozaki, T. Small GTP-binding proteins. *Physiol. Rev.* **81**, 153–208 (2001).

Tanford, C. *The Hydrophobic Effect: Formation of Micelles and Biological Membranes.* John Wiley & Sons, New York, 1980.

Uetz, P., et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).





从20世纪40年代起，Max Perutz和John Kendrew就意识到解析蛋白质结构的意义，他们解析了血球蛋白和肌球蛋白的结构。因为这项工作，他们共同获得了1962年的诺贝尔化学奖。上图为肌凝蛋白晶体的X射线衍射图（参考 http://www.nobel.se/chemistry/laureates/1962/kendrew-lecture.pdf)，Kendrew研究的是抹香鲸（*Physeter catodon*）的肌球蛋白，他用同晶置换的方法将一个重金属离子引入蛋白质晶体中，然后用X射线照射晶体以获得一个X射线衍射图谱（如上图所示）。根据衍射图谱可以推断出晶体的电子云分布，这需要对图谱上的25000个斑点进行分析。下图为Perutz和Kendrew用的EDSAC Ⅰ电脑（1949年引进，参见 http://www.cl.cam.ac.uk/Relics/jpegs/edsac99.36.jpg)。这台电脑对分析X射线的衍射图谱是必需的。我们可以登录 http://www.dcs.warwick.ac.uk/~edsac/，体验一下EDSAC计算机的模拟机器的处理能力

第 9 章

蛋白质结构

9.1 蛋白质结构与结构基因组学概要

一个公认的结论是：蛋白质的三维构象决定其功能，而目前我们获得的高分辨率蛋白质结构大约有 20000 个，远远小于序列的数量。传统上，为了获知结构，研究者们总是从了解蛋白质的功能开始（图 9-1），然而结构基因组方法是基于相反的策略：利用基因组测序工程产生的大量的预测蛋白编码区（参见第 12～第 17 章）。对于研究这些预测蛋白质来说，最重要的一个内容是获得其结构。我们可以先表达预测的蛋白质，然后用传统的实验手段或计算机模拟得到结构信息，最后再对其功能进行研究（图 9-1）。最近已证实的成千上百个预测蛋白质有助于研究者们基于不同的标准对需要解析的结构进行选择（见下文）。

> 两个蛋白质如果具有相似的结构，往往被假定为具有相似的功能。例如，两个受体蛋白具有类似的结构，即使它们与受体结合的能力或在信号转导方面有所区别，但它们最基本的功能还是相近的。

结构基因组学领域的远期目标是解出序列空间中所有序列所对应的结构（Thornton 等，2000；Burley 等，1999；Burley，2000；Koonin 等，2002）。序列空间可以用蛋白质序列家族来定义，蛋白质序列家族实际上是指那些序列相似度大于 30% 的蛋白集合。因此结构基因组学的目标在于从每一个蛋白质序列家族中至少解出一个高分辨率结构。这个目标从某种意义上说与人类基因组计划同样重要，但是代价更高，需要更多的时间。

9.1.1 蛋白质结构，同源和功能基因组学

正如第 8 章所说，蛋白质研究所关心的一个最基本的问题是功能。人们通常根据未知蛋白与功能已知（或可以推测）蛋白的同源性来进行功能预测（Holm，1998；Domingues 等，2000）。

各种各样的 BLAST 搜索可用以确定这种同源关系（第 4 章和第 5 章）。然而对于很多的蛋白质序列来说相似性却极其有限。以视黄醇结合蛋白（retinol-binding protein，RBP）和增味剂结合蛋白（odorant-binding protein，OBP）为例：它们都属于载脂蛋白家族（lipocalin），分子质量大约为 20ku，均为含量丰富的分泌蛋白。它们都具有一个作为载脂蛋白

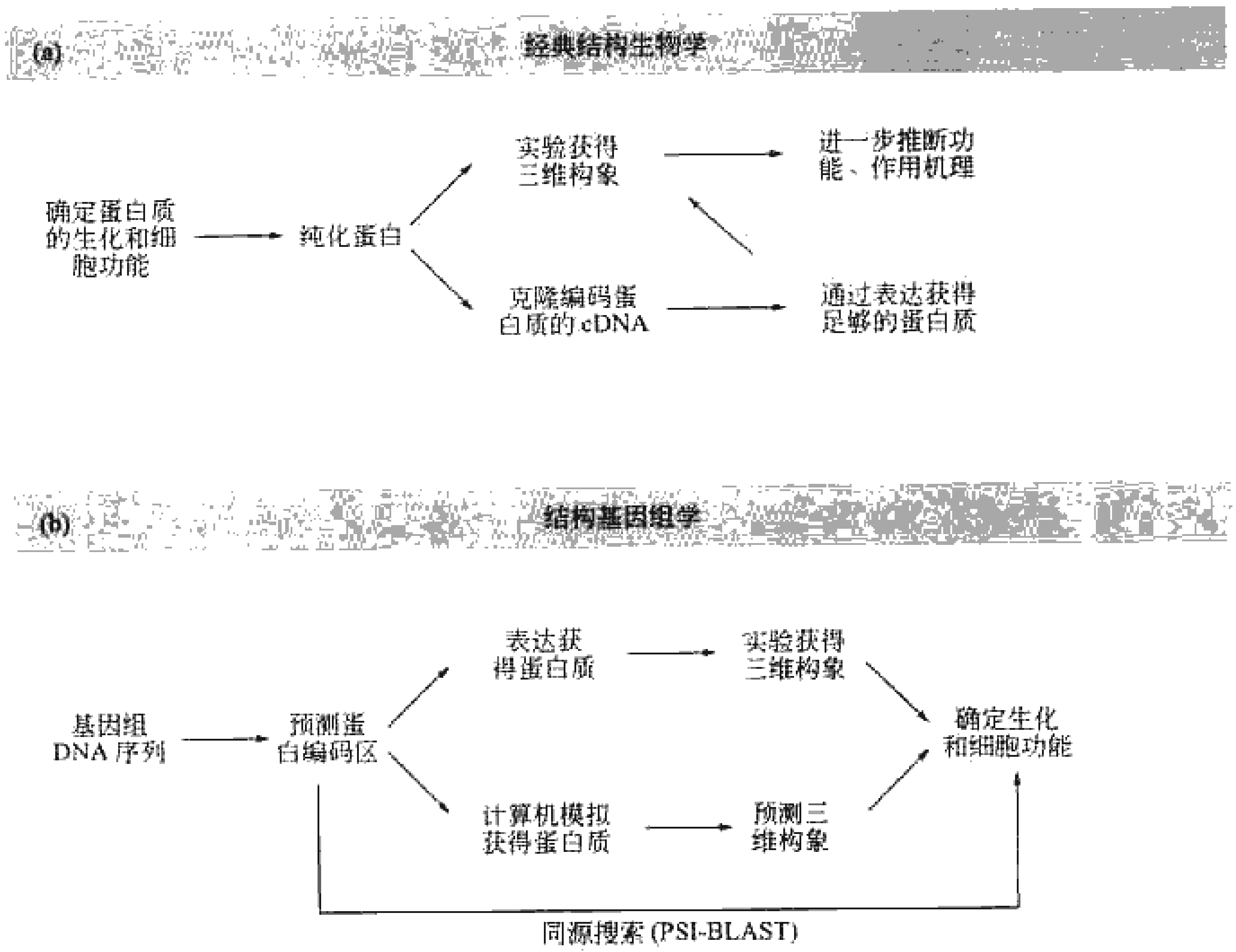


图 9-1 (a) 经典的结构生物学方法。先基于已知的功能或活性对蛋白质进行纯化；经过生化提纯后的蛋白质，如果产量足够，则可以用来结晶进而测得其结构，得到结构后反过来可以进一步研究这种蛋白质的生化特性与作用机理；如果产量不足，则可以将相应的 cDNA 克隆出来（假设已知序列），表达足够的重组蛋白进而纯化以供结构分析之用。(b) 结构基因组方法。从基因组 DNA 出发，预测蛋白编码区，然后通过表达或者计算机模拟的方法得到蛋白质；接着可以用 X 射线晶体衍射或核磁共振等实验手段获得三维构象，同时也可以用计算机进行结构预测；最后基于结构特征可以确定该蛋白质在生化过程中所扮演的角色。另外通过对其序列进行同源搜索（比如用 PSI-BLAST）可以得到关于其生化功能的进一步信息

特征的 GXW 模体（motif），但是通过初级比对分析很难发现它们之间的同源性：两个序列相似度低于 20%。但是这两种蛋白质的三维构象极其相似，可见在进化历程中结构和功能较之序列更为保守。

很难对大鼠（rat）的视黄醇结合蛋白（P04916）与增味剂结合蛋白（NP-6202580）进行两两比对。如果用 BLAST 2 Sequences 工具，你将找不到显著的匹配（即使你使用了较高的 *E* 值和适用于远缘蛋白的 PAM250 打分矩阵）。但是如果用大鼠的 OBP 为输入进行 PSI-BLAST 搜索，你将在多次迭代之后发现视黄醇结合蛋白。

我们能否总结出氨基酸序列相似性与蛋白质结构之间的关系？很显然，一个氨基酸残基的替换可能会引起结构上的巨大变化，例如某些致病的突变（本章结束时会讨论）。很多其他的替换并没有对结构造成很大影响，正如前面载脂蛋白的例子。序列比结构进化更快是一个很普通的现象。Wood 和 Pearson（1999）研究了 36 个蛋白质家族，其中每一个家族中至少有 5 个成员的三维结构已知。他们发现大约 3/4 的蛋白质家族中序列相似性与结构相似性之间有很强的线性关系。因此他们得出结论：大多数氨基酸序列变化能造成可察觉的构象改变，同时此构象变化率在同一个家族内部保持相对恒定。

结构生物学的一个目标是确定所有主要蛋白质家族的三维结构

(Gerstein 和 Levitt，1997；Wolf 等，1999；Koonin 等，2002)。这将有助于我们对蛋白质结构域亲缘关系的深入理解，同时对确定很多蛋白质的功能也有所帮助。

9.1.2 结构基因组学提出的问题：载脂蛋白家族

我们用载脂蛋白家族来阐明结构生物学的几个关键问题。

① 每一个蛋白质转运什么配体？对于大多数情况来说，答案是未知的。能否通过对结构的研究来显示结合结构域，从而确定其配体？我们到底需要多少结构方面的信息才能从序列上来预测配体？

② 一些载脂蛋白基因的突变会引起多种疾病。我们能否预测某个突变带来的结构和功能上的变化？

③ 基于进化树与胞内定位的分析，载脂蛋白可以分为几个亚类。这些亚类在多少程度上反映了结构与功能的相似性？

④ 进行基因组测序时，人们发现了一个可能编码载脂蛋白的基因，我们能否利用其他已知结构的载脂蛋白来预测它的结构？

9.1.3 蛋白质结构的基本原理：从一级结构到二级结构

蛋白质结构是定义在不同层次上的。初级结构指的是多肽链中线性的氨基酸残基序列，每两个氨基酸的 α 碳原子是通过由前一个氨基酸的羧基与后一个的氨基形成肽键连接起来。二级结构诸如 α 螺旋、β 折叠和无规卷曲等是由初级结构按一定的规律折叠组装而成，如标准 α 螺旋为每圈 3.6 个残基，第 n 个残基的 C═O 和第 $n+4$ 个残基的 NH 之间形成氢键；而形成 β 折叠的多肽链充分伸展，各肽键平面折叠成锯齿状结构，侧链 R 基团交错位于锯齿状结构上下方，依靠链间肽键 C═O 和 NH 形成氢键维系构象稳定。三级结构是由结构域在三维空间按一定的方式排列形成，所谓结构域是由二级结构元素组成的球形结构单位。最后的四级结构则是由几条具有三级结构的肽链组成。一个蛋白质功能上的重要区域（如配体结合位点或酶的活性区域）是在三级结构或四级结构层次上形成的。

在自然界中，氨基酸序列将决定蛋白质的三维构象。这种构象可能会依赖于一些翻译后修饰，如糖基化与二硫键形成。但是每个细胞都能从初级结构中获得修饰信息，从而形成合适的三维结构。结构生物学家面对的挑战是，我们对仅基于序列数据预测三维结构的方法只有很模糊的认识。

二级结构是由氨基酸侧链的构象决定的。肌红蛋白基本由 α 螺旋组成（图 9-2）。典型的 α 螺旋长度在 4～40 个氨基酸残基。β 折叠结构是由蛋白质分子中舒展的 β 链组成，长度大约在 5～10 个残基，β 链以平行或反平行的方向排列。一个典型的例子是来自 *E. coli* 的硫氧化还原蛋白（图 9-2）。大多数蛋白质同时具有 α 螺旋和 β 折叠。

相对于三级结构，二级结构预测要容易一些。20 世纪 70 年代，Chou 和 Fasman（1978）发展了一种预测二级结构的方法，其原理是：不同的残基在 α 螺旋、β 折叠以及无规卷曲中出现的频率有所差异，例如脯氨酸几乎不会发生在 α 螺旋中，它往往出现在转角的位置。他们先计算各种残基在一定的上下文（沿着序列移动的窗口）中形成 α 螺旋、β 折叠以及无规卷曲的倾向性，然后沿着蛋白质序列扫描，找出连续 6 个残基中至少有 4 个 α 螺旋或其他二级结构类型（β 折叠、卷曲和转角等）得分高于阈值的区域，扫描可以向任意一个方向延伸。

以后又发展了很多方法，例如 Garnier，Osguthorpe 和 Robson（1978）的 GOR 方法

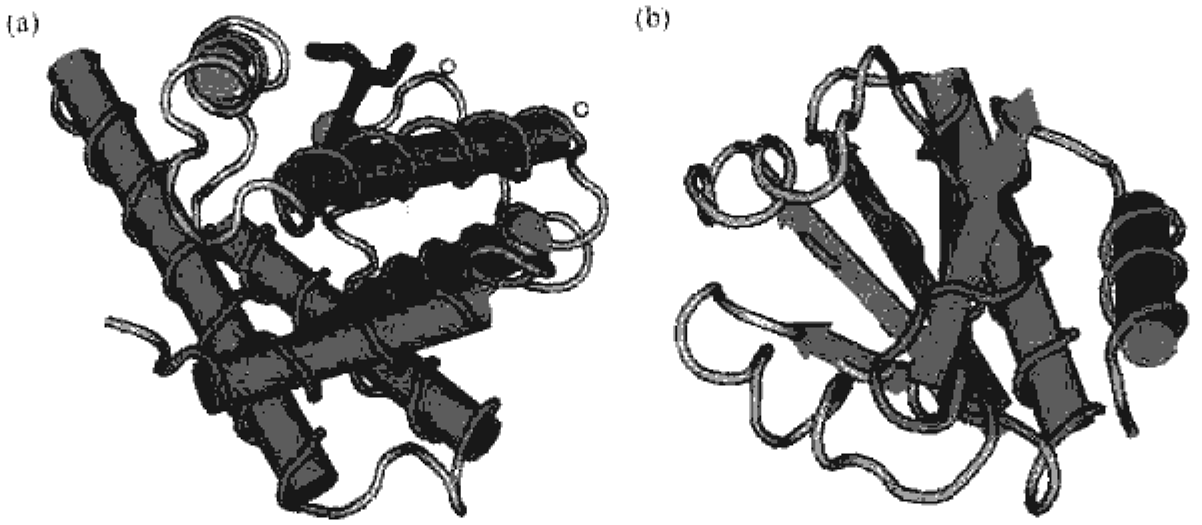


图 9-2 二级结构的例子。(a) 肌红蛋白 (myoglobin, PDB标志码为4MBN) 由大片区域的α螺旋组成 (图中显示为包在桶状物周围的绳索状结构); (b) 大肠杆菌中的硫氧化还原蛋白 (thioredoxin, PDB标志码为1TXX) 含有5束β折叠 (图中显示为5个箭头)。对蛋白质而言最普遍的情况是同时具有α螺旋与β折叠。蛋白质分类学在很多数据库中均有描述，如CATH (见下文)

(Garnier等，1996)。在很多情况下这些算法被用来分析单个序列 (直到现在，它们仍然很有用)。由于现在可以很容易得到序列的多重比对，二级结构预测的精度也得到了提高。PHD程序 (Rost和Sander，1993 a，1993 b) 就是一个用多序列比对来进行二级结构预测的优秀程序。将各种算法作用于已知结构的序列，根据结果可以对这些算法进行评估。目前，一般可以达到70%～75%的精度 (Rost，2001；Przybylski和Rost，2002)，这已经远远超过了Chou-Fasman算法的水平。

PBIL 网址 http://npsa-pbil.ibcp.fr

有很多网络服务器都提供预测二级结构服务 (表9-1)，有些只允许提交一条序列，而一些却可以接受多重序列比对。PBIL的服务器提供9种算法得到的二级结构预测 [图9-3(a)]，各种预测方法的结果有细微差异，但大体上是相似的。

表 9-1 网上的二级结构预测程序

程序	评注	网站地址
APSSP	基于神经网络方法	http://imtech.ernet.in/raghava/apssp/
GOR4	来自波兰的Bio-Informatique Lyonnais	http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl? page=npsa_gor4.html
Jpred	来自Barton小组	http://www.compbio.dundee.ac.uk/~www-jpred/
NNPREDICT	加强的神经网络方法	http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html
PHD	基于神经网络方法	http://cubic.bioc.columbia.edu/predictprotein
Predator	来自Argos小组	http://www-db.embl-heidelberg.de/jss/servlet/de.embl.bk.wwwTools.GroupLeftEMBL/argos/predator/predator_info.html
PredictProtein server	来自哥伦比亚大学	http://cubic.bioc.columbia.edu/predictprotein/
PSIPRED	来自伦敦大学	http://bioinf.cs.ucl.ac.uk/psipred/
SAM-T99sec	利用隐马模型(第10章)	http://www.cse.ucsc.edu/research/compbio/HMM-apps/T99-query.html
Sosui	来自Mitaku小组	http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html

注：更多站点列表见ExPASy (http://kr.expasy.org/tools/#secondary)、PBIL (http://npsa-pbil.ibcp.fr) 和EVA (http://cubic.bioc.columbia.edu/eva/)。

(a)

```
                     10         20         30         40         50         60         70
                      |          |          |          |          |          |          |
UNK_266410  AQEEEAEQNLSELSGPWRTVYIGSTNPEKIQENGPFRTYFRELVFDDEKGTVDFYFSVKRDGKWKNVHVK
DPM         cchhhhhhchcctctcceeeeeettctccchtcctccccceehhhehhthccceeeeeeettttcctcceeeh
DSC         cccccccccceeeeccceeeeeeeccccceeeeecccccceeeeeeeeeccccceeeeeeeeeccccceeeeeee
GOR4        ccccccchhhhhhccccceeeeecccccccccccccccccceeeeeeeccccccceeeeeeeeeccccccceeeee
HNNC        cchhhhhhhhhhhcccceeeeeeecccccccccccccceheehheeeccccccceeeeeeeeccccccceeeeeee
PHD         cccccccccccccceeeeeeeeeecccccccccccceeeeeeeeeeeecccccceeeeeeeeeeeccccceeeeeee
Predator    ccchhhhhhhcccccceeeeeeccccccccccccccchhhhhhhcccccceeeeeeeccccccceeeeeee
SIMPA96     chhhhhhhhhhhcccccceeeeeecccccccccccchhhhhhheeccccccceeeeeeeeccccccceeeeee
SOPM        hhhhhhhhhhhhhtccceeeeeeeeeccttcccttcccchhhhheeecttcceeeeeeeeettcccceeee
Sec.Cons.   cc?hhhhhhhhcccceeeeeeccccccccccccccc??e?heeccccccceeeeeeeeccccccc?eeeee

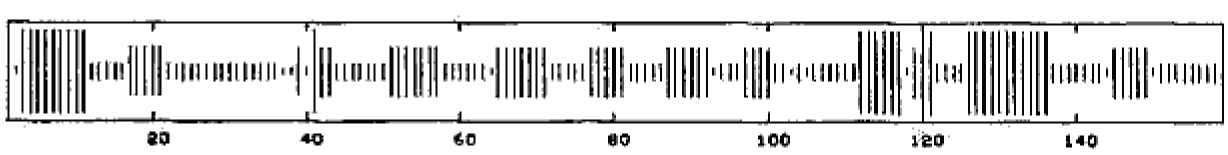
                     80         90        100        110        120        130        140
                      |          |          |          |          |          |          |
UNK_266410  ATKQDDGTYVADYEGQNVFKIVSLSRTHLVAHNINVDKHGQTTELTELFVKLNVEDEDLEKFWKLTEDKG
DPM         hhttttccccehctcttcceeeeeeeeeeeheehhcetetcccccchhhhheehhhhhhhhhhhhhhhhtcc
DSC         eecccccceeeeeccccceeeeeeccceeeeeeeeeeeecccccceeeeeeeeeeccccccchhhhhheecccc
GOR4        eeccccceeeeeeccccceeeeeeeccchhhhhcccccccccccchhhhhhhhhhhhcchhhhhhhhhhhhccc
HNNC        eeccccceeeeeeccccceeeeeeeeccceeeeecccccccccccchhhhhhheeeecccchhhhhhhhhhhccc
PHD         eeecccceeeeeeeeeeeeeeeeeccceeeeeeeeeeecccccceeeeeeeeeeccccccchhhhhhhhhhhhhcc
Predator    eeecccccceeeeecccccceeeeeeecccccccccccccccccchhhhhhhhhhhcccchhhhhhhhhhhccc
SIMPA96     cccccccceeeeeccccccceeeeeeccccceeeeecccccccccccchhhhhhhhhhhcccchhhhhhhhhhhccc
SOPM        eccttttceeeeeettcceeeeeeeeccceeeeeeeeeectcccchhhhhheeeeeccccthhhhhhhhhhcctt
Sec.Cons.   eecccccceeeeeeccccceeeeeee?cccceeeecc?c?cccccchhhhhh?ehhccccchhhhhhhhhhhhccc

                    150
                     |
UNK_266410  IDKKNVVNFLENEDHPHPE
DPM         ctccccceeehhhhtcctcccc
DSC         cceeeeeeeeeeecccccccc
GOR4        ccceeeeeeeeecccccceeec
HNNC        ccccceeeeeeeeccccccccc
PHD         ccccchhhceeccccccccc
Predator    cccccccceeeccccccccccc
SIMPA96     cchhhhhhhhhhccccccccc
SOPM        cchhhhhhhhhhhctcccccc
Sec.Cons.   cccc?eeeee?cccccccc


 Sequence length :    159
```

(b)

```
Sec.Cons. :
   Alpha helix        (Hh) :    29 is   18.24%
   3_10 helix         (Gg) :     0 is    0.00%
   Pi helix           (Ii) :     0 is    0.00%
   Beta bridge        (Bb) :     0 is    0.00%
   Extended strand    (Ee) :    43 is   27.04%
   Beta turn          (Tt) :     0 is    0.00%
   Bend region        (Ss) :     0 is    0.00%
   Random coil        (Cc) :    76 is   47.80%
   Ambigous states    (?)  :    11 is    6.92%
   Other states            :     0 is    0.00%
```




```
PREDATOR parameters :
   Secondary structure data : dssp

SOPM parameters :
   Window width          :  17
   Similarity threshold  :   8
   Number of states      :   4
```

View prediction result file in MPSA : [DPM] [DSC] [GOR4] [HNN] [PHD] [PREDA] [SIMPA96] [SOPM]
View prediction result file in ANTHEPROT : [DPM] [DSC] [GOR4] [HNN] [PHD] [PREDA] [SIMPA96] [SOPM]

图 9-3 几种提供二级结构预测的网络服务器。提交一种载脂蛋白的序列（牛的增味剂结合蛋白，bovine odorant-binding protein，获取码 P07435）到波兰 Bio-Informatique Lyonnais 的蛋白质序列分析网络（Network Protein Sequence Analysis）的网站。(a) 和 (b) 为联合了 9 种算法得到的二级结构预测结果。请注意这些算法提供的结果稍有区别

9.1.4 蛋白质三级结构：蛋白质折叠问题

蛋白质是如何折叠成三维构象的？这个过程在自然情况下能很快完成。通常是在合成完毕时，折叠也差不多完成了。有一些蛋白质只有在它们穿越内质网脂双层后，在内质网的内腔中才发生折叠。

在结构生物学中，主要有3种确定蛋白质结构的方法。

① 通过实验的方法，主要是X射线晶体学和核磁共振的技术。

② 比较-同源模建方法。通过将目标蛋白序列与一个或几个已知结构的同源蛋白（模板）进行比较，从而预测目标的结构（Marti-Renom等，2000；Shortle，2000）。当目标与模板的序列相似度较高时比较-同源模建相当有效。

③ 从头预测的方法。这种方法仅通过物理学原理来预测目标的结构（Osguthorpe，2000；Simons等，2001）。

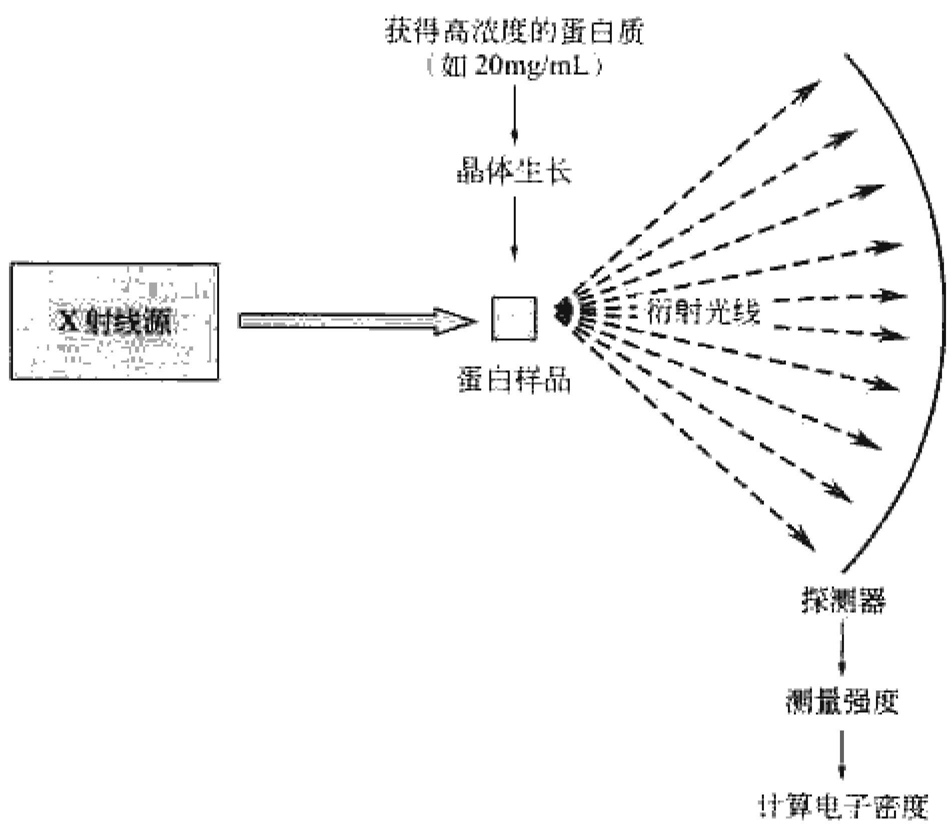
在本章中我们将先简要介绍如何用实验手段确定蛋白质结构，接着介绍PDB数据库，然后是如何利用NCBI、ExPASy和其他的工具，以使结构可视化。还有很多关于蛋白质分类的数据库，如CATH，SCOP和Dali。讨论完这些数据库后，我们将研究两类预测蛋白质结构的计算方法：比较-同源模建和从头预测方法。

9.1.5 获得蛋白质结构的实验手段

目前主要有两种方法可以得到蛋白质的结构：X射线衍射和核磁共振技术。X射线衍射技术是用于测定蛋白质结构最精确的方法（框9-1），目前已知结构中的80%是通过这

框9-1 X射线晶体学

首先需要获得高浓度的蛋白质，然后在硫酸铵之类的溶剂内结晶，将一束X射线瞄准蛋白质晶体。高度有序的蛋白质阵列使X射线发生衍射，被感光胶片探测到。测量点的亮度，利用Fourier变换重构图像，同时产生对应于蛋白质原子排列的电子密度图。单个原子之间距离为0.1～0.15nm，如果分辨率低于0.2nm则需要更精细的测定。



种技术得到。蛋白质溶液样品必须有足够高的浓度并在合适的条件下形成晶体。晶体将X光衍射到探测装置，晶体结构可以通过衍射图像类型推导得到。X射线的波长（0.05～0.15nm）适用于测量原子间距，因而可以用来对蛋白质侧链进行细致的研究。最早解出结构的蛋白质是肌红蛋白（myoglobin，John Kendrew，1958）[图9-2(a)]，这种高度不规则的形态反映了氨基酸残基复杂的空间组织。总体上说蛋白质倾向于将疏水性残基包裹在内部，亲水性残基暴露在表面。虽然蛋白质的主链是高度极性的，这样的疏水核心还是可以形成——蛋白质通常将内部的残基组织成二级结构来降低主链极性。虽然我们可以对二级结构特征进行预测，但是想预测这些二级结构是如何折叠成为三维构象却相当困难。

核磁共振谱是另一种重要的测定蛋白质结构的方法。将蛋白质溶液置于磁场中，观察特征性的化学位移，从化学位移中推测蛋白结构。由核磁共振谱测定的蛋白质最大约含有350个残基（约40ku），这比由X射线晶体衍射测得结构的蛋白质小得多。但是核磁共振谱有一个优势：不需要复杂的结晶过程。

9.1.6 选择目标与获得蛋白质的三维构象

从选择目标开始（Brenner，2000），用实验方法获得蛋白质结构的一般流程已在图9-4中列出。以前，通常选择血红蛋白、细胞色素C类便于实验操作的蛋白质：它们体积小、水溶性好、含量丰富，而且具有重要的生物学功能。如今在选择目标方面要考虑以下一些新的标准（McGuffin和Jones，2002）。

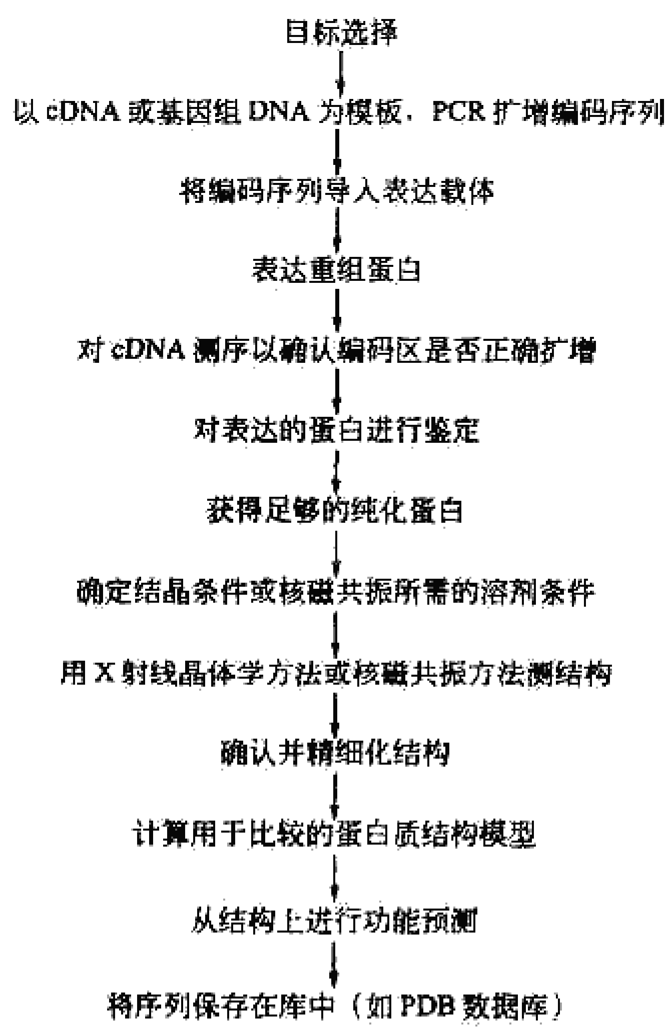


图9-4 获得蛋白质三维结构的大致流程（来自Burley，2000，有改动）

① 要研究生命的所有分支（真核生物、细菌、古细菌和病毒）。

② 是否应该倾向于从以前没有研究过的蛋白家族中选择成员？

③ 医学上有重要意义的蛋白质（如药物靶标）是否应该优先考虑？

④ 是否应该投入大量精力把某一个体的所有蛋白质结构的全部测定？

⑤ 如何测定具有跨膜结构域的蛋白质？这些蛋白的研究在技术上具有相当的难度。Chang和Roth（2001）成功解出了*E. coli*的一个多药物抗性ABC转运蛋白，他们共尝试了96000种结晶条件，终于找到了一些适用于X射线衍射测定。

9.2 PDB数据库

PDB（protein data bank）数据库负责收集蛋白质结构信息，其中收录了大量蛋白质的三维结构文件（Westbrook等，2002；Berman等，2002），记录有原始结构数据：原子坐标、配基的化学结构和晶体结构的描述等。PDB通过评估模型质量和它们与实验数据的吻合程度来证实结构。

> PDB于1971年建于位于美国长岛的Brookhaven国家实验室，于1998年转移到结构生物信息学研究合作组织（Research Collaboratory for Structural Bioinformatics，RCSB），网址：http://www.rcsb.org/pdb/。

PDB网站主页见图9-5。这个数据库目前拥有超过20000个的结构记录（表9-2），而且数据增长速度相当快（图9-6）。通过获取码（accession number）可以访问该数据库，PDB获取码由1个数字和3个字母组成（如血红蛋白的获取码为4HHB），同时PDB数据库也支持关键词搜索。对载脂蛋白进行关键词搜索的结果见图9-7。点击牛增味剂结合蛋白（bovine odorant-binding protein）的“explorer”链接，打开结构查看（structure explorer）页面（图9-8）。列表信息包括分辨率（resolution）、空间群（space group）和晶体的单位晶胞尺度（unit cell dimension）。“View Structure”选项链接到一系列结构可视化工具（图9-9），常用的一些列于表9-3。在本章后面，我们将对其中一部分进行深入的研究。



图9-5 PDB（protein data bank）是蛋白质与其他大分子结构的仓库（Berman等，2000）（http://www.rcsb.org/pdb），可以用多种方法进行查询

表 9-2 PDB 内容列表（2003 年 6 月）

实验技术	分子类型				
	蛋白质多肽病毒	蛋白-核酸复合体	核 酸	糖 类	总 数
X 射线衍射与其他	16452	788	673	14	17927
核磁共振谱	2609	91	520	4	3224
总数	19061	879	1193	18	21151

注：来源于 http://www.rcsb.org/pdb/holdings.html。

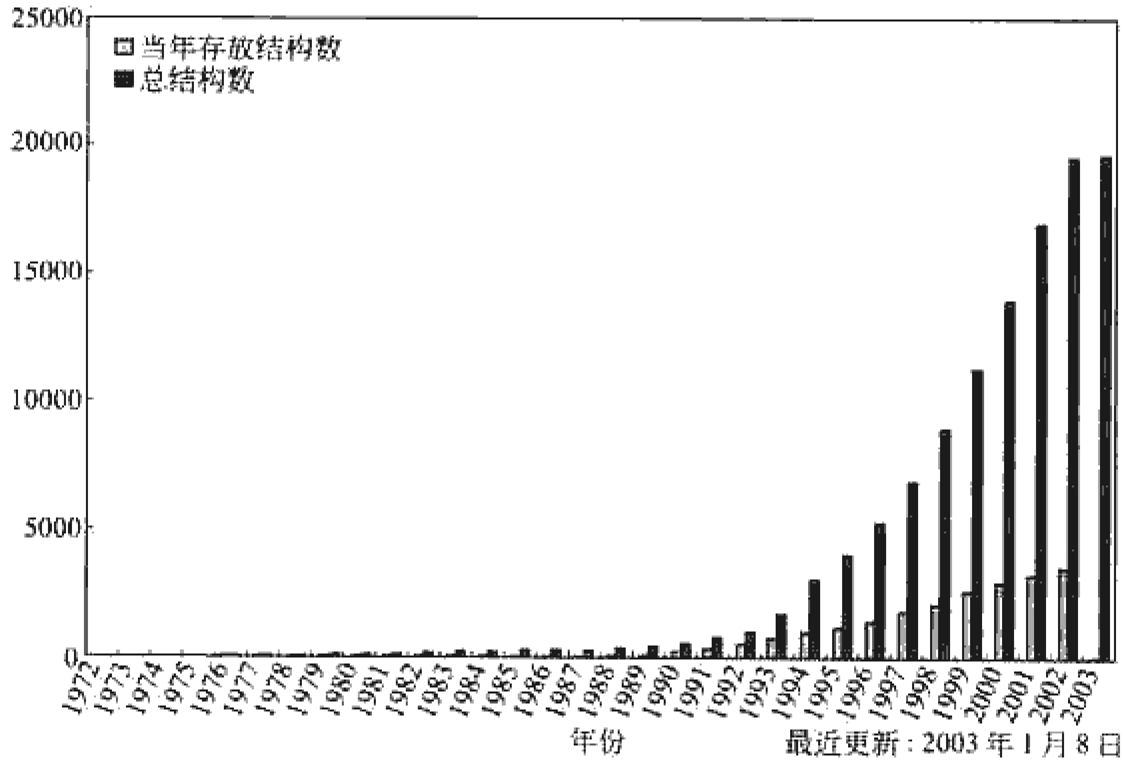


图 9-6 PDB 数据库规模在过去的 10 年中猛增。目前这个数据库的概况可以参照 http://rcsb.org/pdb/holdings.html

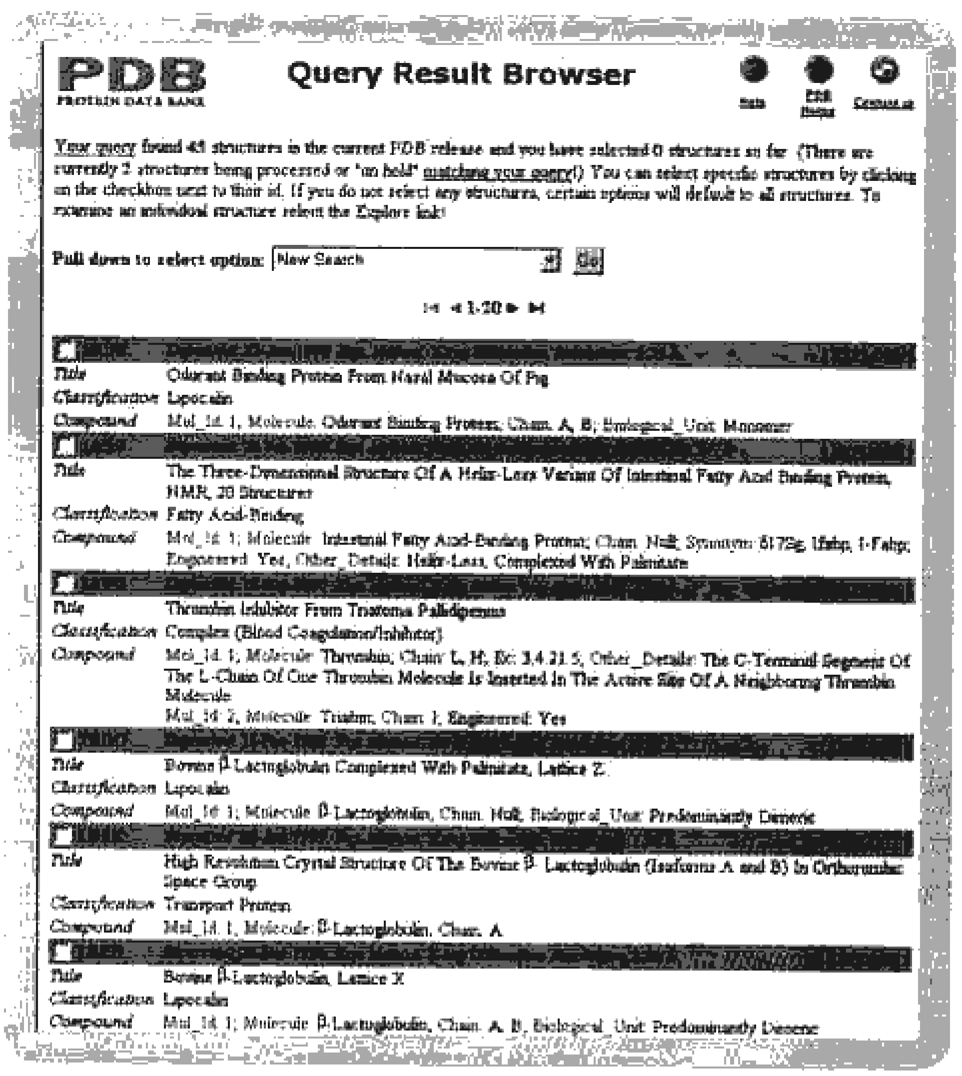


图 9-7 在 PDB 查询载脂蛋白的结果

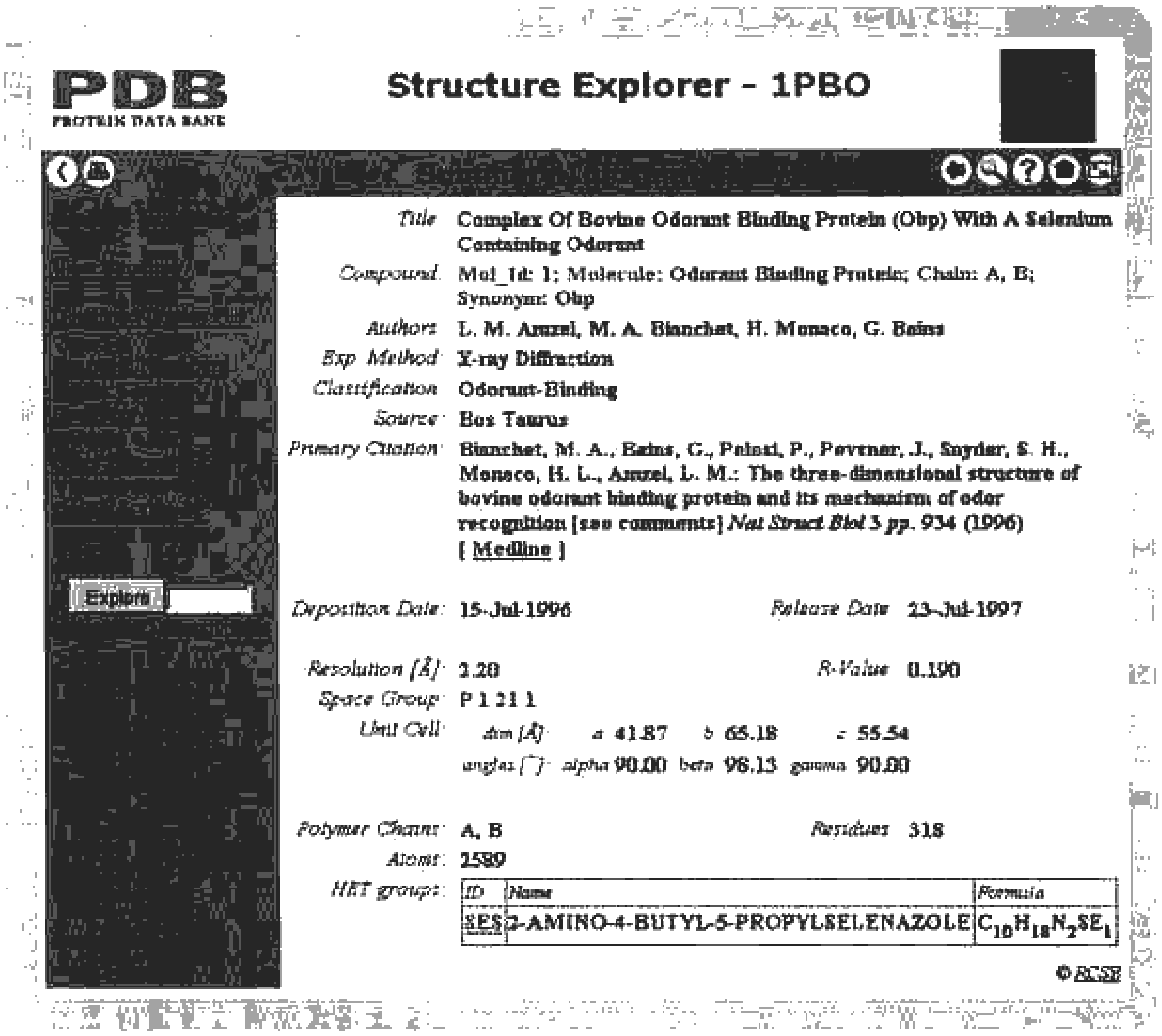


图 9-8 搜索增味剂结合蛋白（odorant-binding protein，OBP，载脂蛋白的一种）结构的结果。概要信息（summary information）包括了对分辨率（resolution）、空间群（space group）、单位晶胞尺度（unit cell dimension）等的描述。这个蛋白是在硒类衍生物存在的情况下结晶的。左边一列包括了结构观察（view structure）等其他信息的链接

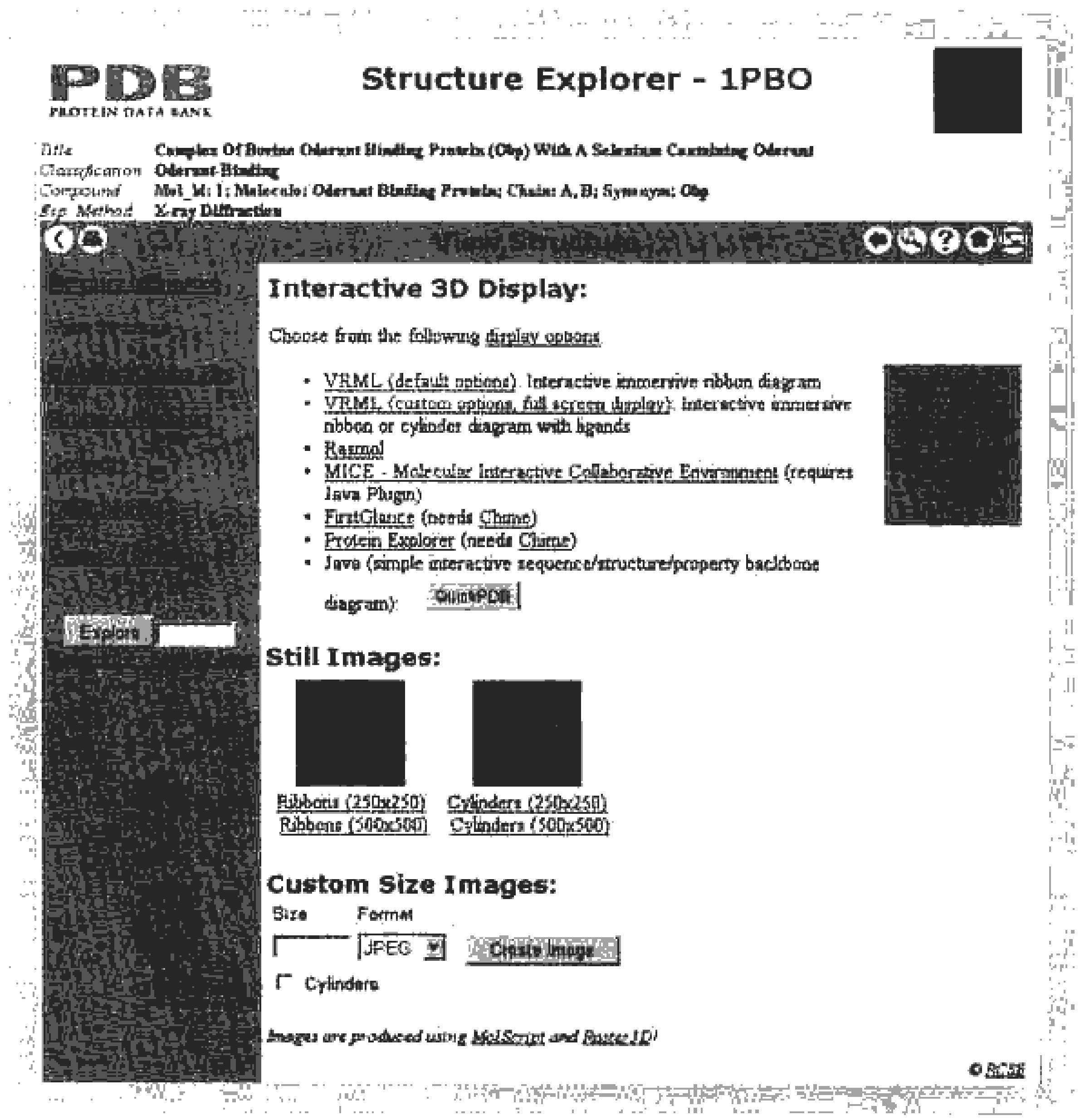


图 9-9 PDB的结构观察（View Structure）页面提供了OBP结构的静态图片和很多显示选项：VRML、Rasmol、MICE、FirstGlance和Protein Explorer

表 9-3 用于结构观察的可视化工具

工具	评注	网站地址
Chime	浏览器插件	参照 PDB 的说明
Cn3D	来自 NCBI	http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
Mage	读取 Kinemages 文件	http://www.ncbi.nlm.nih.gov/Structure/CN3D/mage.html
MICE Java applet		参照 PDB 的说明
RasMol	一个单机版软件	参照 PDB 的说明
SwissPDB viewer	位于 ExPASy	http://www.expasy.org/spdbv/
VMD	可视化分子模拟软件,属于伊利诺斯大学	http://www.ks.uiuc.edu/Research/vmd/
VRML	使用 MolScript	参照 PDB 的说明

注：PDB 的 URLs 可以参照帮助文档（htttp://www.rcsb.org/pdb/help-graphics.html）。

还可以在 PDB 站点内用 Pearson 和 Lipman 的 FASTA 程序（1998）进行搜索。在 PDB 主页，选择"SearchFields"（图 9-5 箭头 1 所指）。这个搜索页面包含了一个进行 FASTA 搜索的选项，可以很方便地使用查询访问 PDB 的结构数据。对于较短的查询，必须使用命令来提高 E 值，如"-E10."。

大分子结构数据库（The Macromolecular Structure Database）位于 http://www.ebi.ac.uk/msd/。

除了 PDB，欧洲生物信息学研究所维护的大分子结构数据库（Macromolecular Structure Database）也相当有名。这个数据库与 PDB 整合，它是欧洲生物大分子数据中心。

NCBI 的结构（Structure）页面位于 http://www.ncbi.nlm.nih.gov/Structure/。

PDB 数据库在结构生物学中占有中心地位（图 9-10）。很多数据库和网络服务器都与它链接或者将它的数据直接整合到本地资源里（见网络资源，表 9-6）。我们将按照图 9-10 的顺序来介绍这些资源：首先我们研究 NCBI 和其他一些可以分析单个蛋白质结构或对几个结构进行比较的站点，然后对关于蛋白质比较分类系统的数据库（如 SCOP，CATH 和 Dali 等）进行介绍。

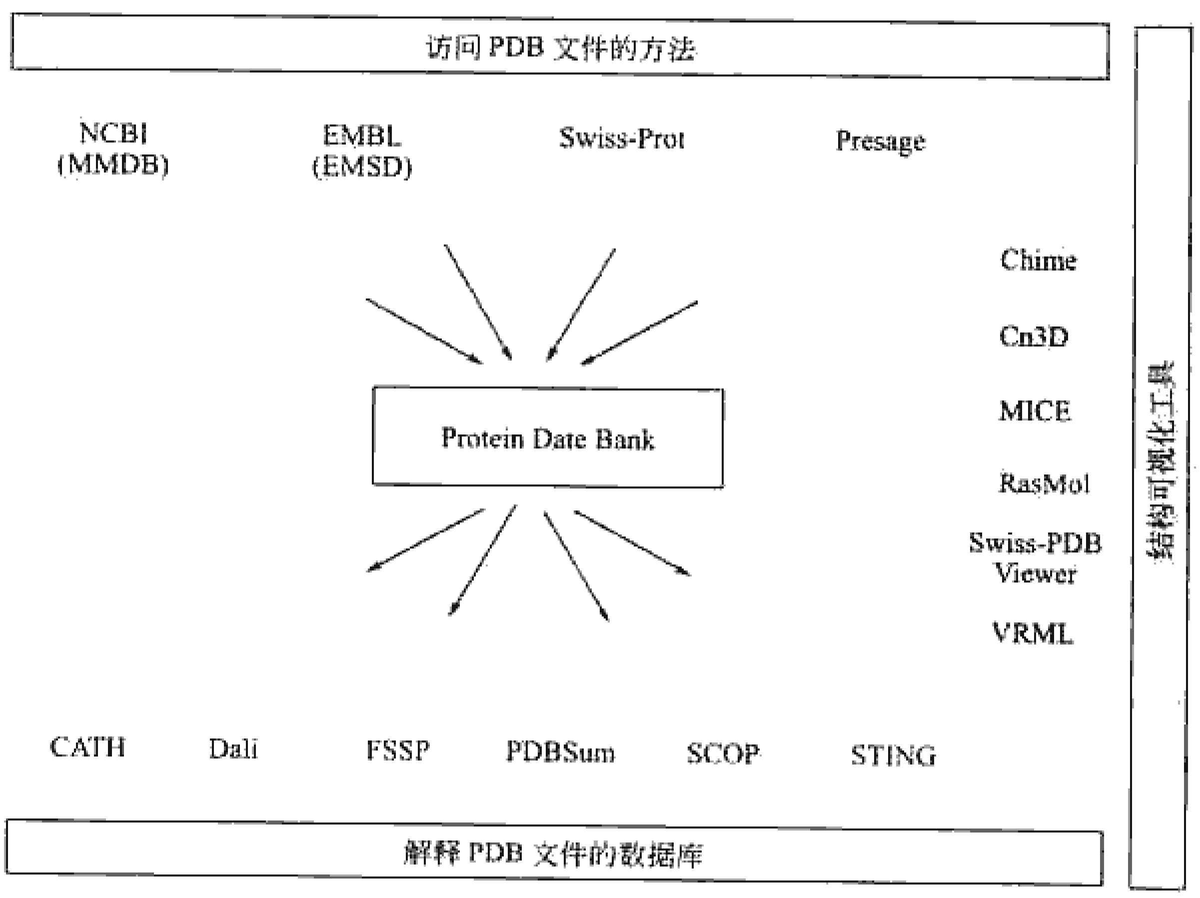


图 9-10 PDB 是存储蛋白质和其他生物大分子结构数据的仓库（http://www.rcsb.org/pdb/.）。可以通过很多站点访问 PDB 数据，如 NCBI 和 EMBL。另外还有很多基于 PDB 数据分析而产生的蛋白质分类数据库，如 SCOP、CATH、Dali 和 FSSP，其他数据库在网络资源中有介绍（表 9-6）

9.2.1 在NCBI访问PDB记录

有很多方法可以在NCBI数据库中获得蛋白质结构信息。

① 可以在Structure页面（图9-11）或通过Entrez用关键词搜索得到PDB结构。

② 利用蛋白质相似性进行搜索。在Entrez蛋白质数据库中选择一个感兴趣的蛋白质，然后找“Related Structures”链接。另一种方法是进行blastp搜索，然后将输出结果限制在PDB数据库内（图9-12）。

③ 搜索核酸数据库。可以对一个DNA序列进行blastx搜索，然后将输出结果限制在PDB数据库内。

在Entrez的结构中查找增味剂结合蛋白（odorant-binding protein），我们可以得到一个PDB标志号（获取码）的列表（图9-13）。点击“3D Domains”可以得到同一个蛋白质不同链的记录（图9-14）。而点击“1PBO”则显示MMDB中关于这个蛋白质的记录（图9-15）。这些就是从NCBI得到蛋白质结构的主要入口（Wang等，2000），包括了文献、分类学数据、序列邻居、结构邻居和可视化选项。

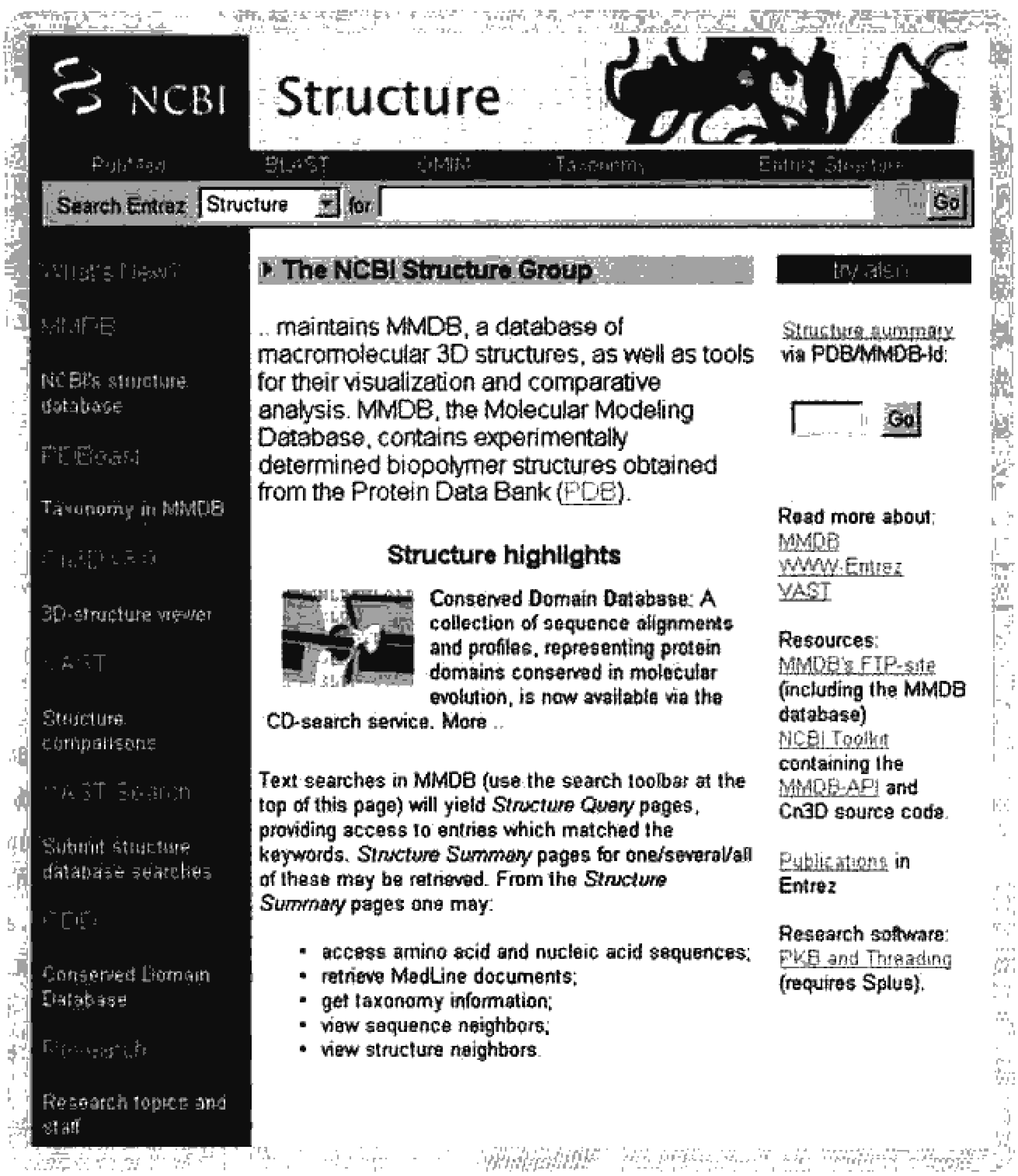
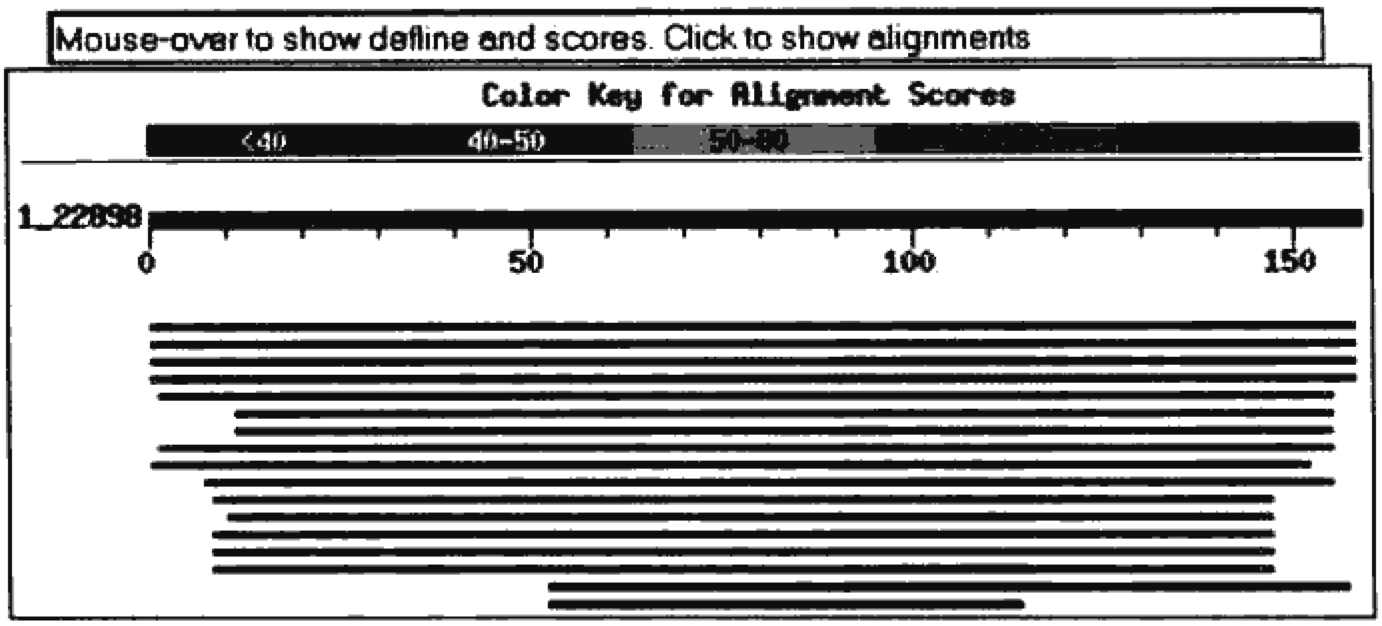


图9-11 NCBI的结构页面提供指向PDB、MMDB和一些工具（如Cn3D结构查看器、VAST结构比对工具）的链接（http://www.ncbi.nlm.gov/Structure/）

Distribution of 17 Blast Hits on the Query Sequence



Related Structures

Sequences producing significant alignments:		Score (bits)	E Value
gi|1827612|pdb|1OBP|A	Chain A, Odorant-Binding Protein From...	293	2e-80
gi|17943132|pdb|1HN2|A	Chain A, Crystal Structure Of Bovine...	292	4e-80
gi|21730166|pdb|1G85|A	Chain A, Crystal Structure Of Bovine...	291	5e-80
gi|2392495|pdb|1PBO|A	Chain A, Complex Of Bovine Odorant Bi...	291	8e-80
gi|12084609|pdb|1DZJ|A	Chain A, Porcine Odorant Binding Pro...	115	5e-27
gi|4558087|pdb|1A3Y|B	Chain B, Odorant Binding Protein From...	111	1e-25
gi|4558086|pdb|1A3Y|A	Chain A, Odorant Binding Protein From...	110	2e-25
gi|5107505|pdb|1BJ7|	Bovine Lipocalin Allergen Bos D 2	77	2e-15
gi|5822446|pdb|2A2G|A	Chain A, The Crystal Structures Of A2...	72	1e-13
gi|15826041|pdb|1E5P|A	Chain A, Crystal Structure Of Aphrod...	70	5e-13
gi|7766863|pdb|1EW3|A	Chain A, Crystal Structure Of The Maj...	58	1e-09
gi|21465464|pdb|1GM6|A	Chain A, 3-D Structure Of A Salivary...	58	1e-09
gi|13399976|pdb|1I04|A	Chain A, Crystal Structure Of Mouse ...	56	7e-09
gi|494384|pdb|1MUP|	Major Urinary Protein Complex With 2-(...	55	9e-09
gi|8569601|pdb|1DF3|A	Chain A, Solution Structure Of A Reco...	55	1e-08
gi|4558223|pdb|2SHP|A	Chain A, Tyrosine Phosphatase Shp-2 >...	28	1.3
gi|640187|pdb|1AYA|A	Chain A, Tyrosine Phosphatase Syp (N-T...	28	1.8

图 9-12 blastp 或 blastx 搜索结果中来自 PDB 数据库的部分，这里搜索牛的 OBP（P07435），得到的结果中包含了多种已知结构的载脂蛋白，有牛的载脂蛋白过敏原 Bos D 2，α2u 球蛋白和来自仓鼠的性外激素结合蛋白 aphrodisin

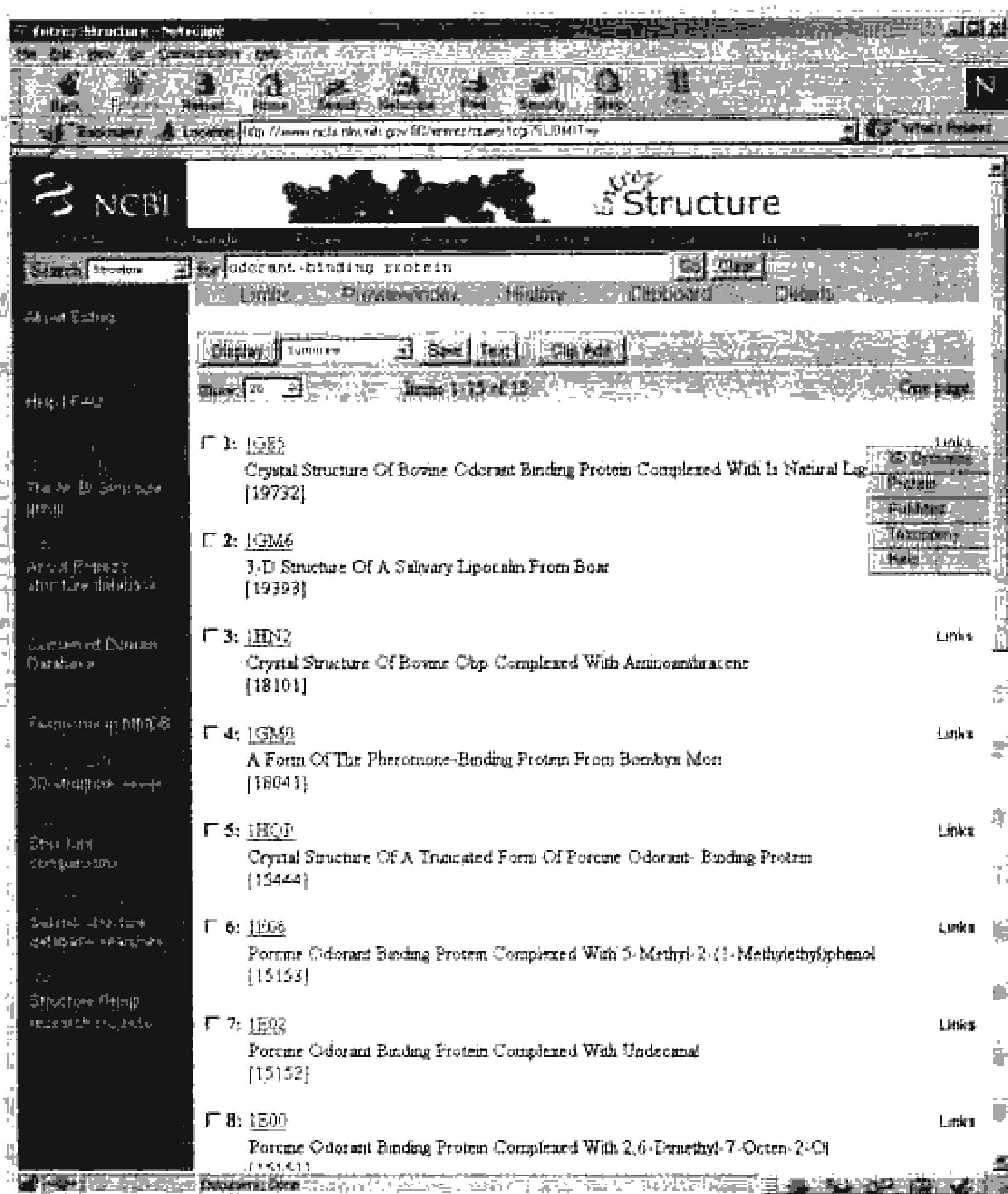


图 9-13 用关键字"odorant"在 NCBI 的结构(structure)页面进行搜索得到的结果。点击结果列表中的 bovine OBP (PDB 获取码 1PBO),结果显示于图 9-15

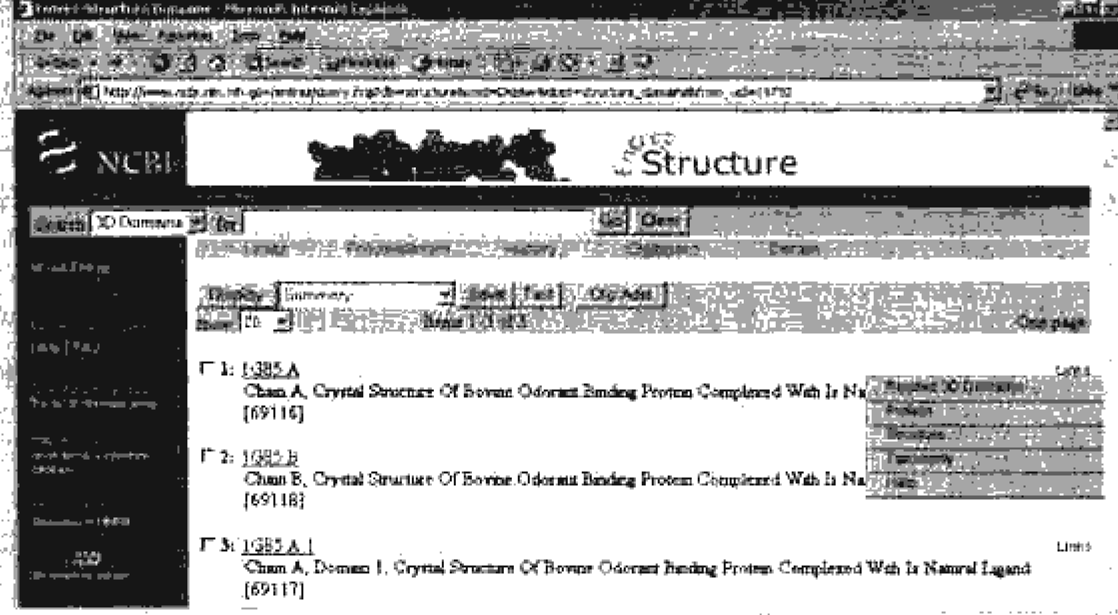


图 9-14 点击"3D domains"的链接(见图 9-13)可以获得 PDB 中同一个蛋白质不同链的记录。而"Related 3D Domain"链接到上百个关于载脂蛋白结构数据库的记录列表

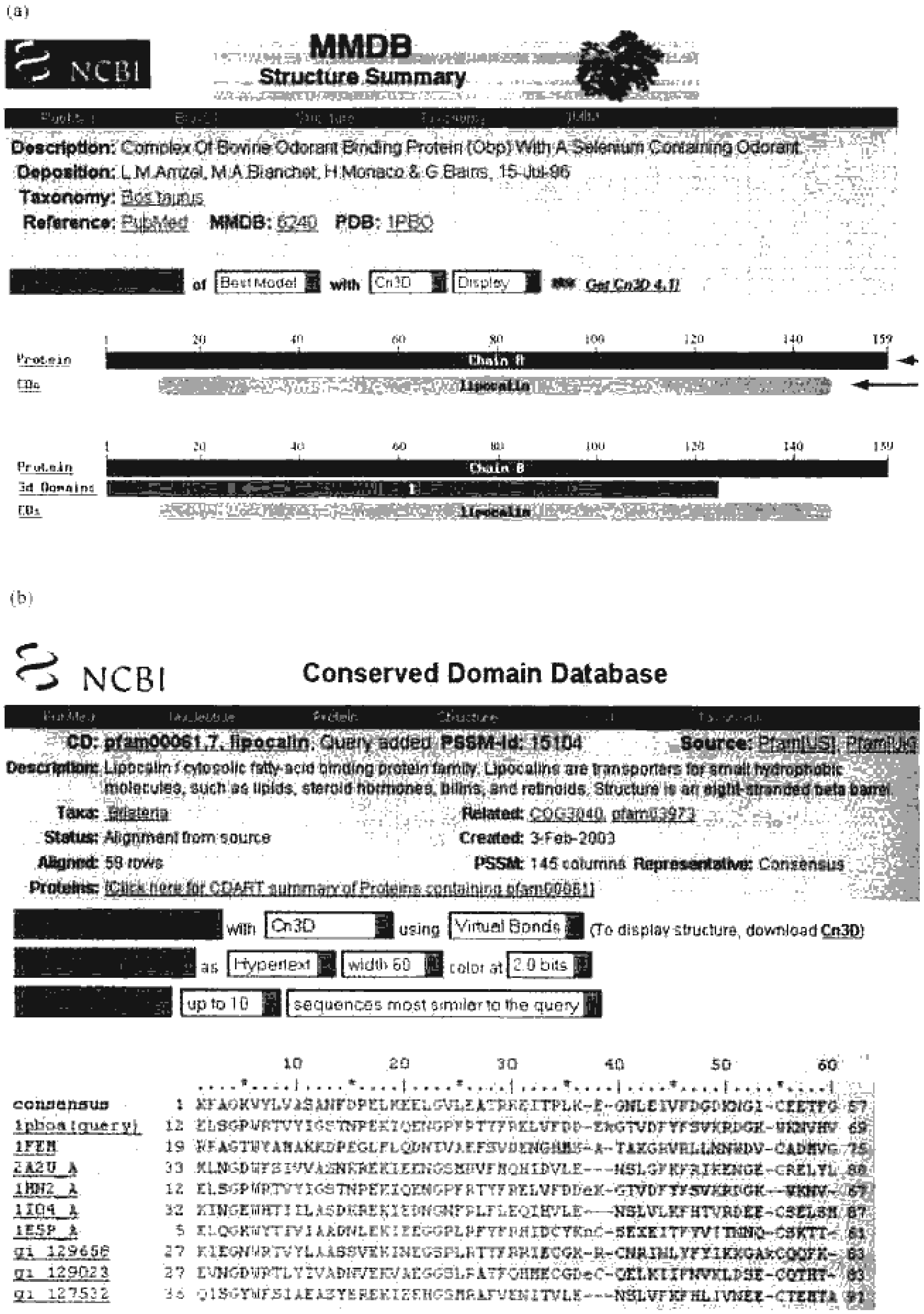


图 9-15 (a) 在 Entrez 蛋白质结构列表（如图 9-13，图 9-14）中点击“1PBO”链接，可以获得 NCBI 分子模拟数据库（molecular modeling database，MMDB）的结构概要。点击“View 3D Structure”链接，可以运行 Cn3D 或其他一些结构查看器（RasMol，MAGE）。另外还有指向邻近结构（structure neighbors）和保守结构域数据库（conserved domain database，CDD）的链接。CDD 链接会提供多序列比对结果，见（b）

Cn3D是see in 3D的首字符缩写。

在MMDB中通过"View Structure"链接可以很容易地观察蛋白质结构。下载Cn3D后会打开两个窗口：Cn3D Viewer和OneD-Viewer（图9-16）。Cn3D Viewer可以用7种格式来显示蛋白质结构，同时可以对结构进行旋转。相应的OneD-Viewer则显示此蛋白质的氨基酸序列，包括α螺旋与β片层的区域。在Cn3D Viewer或OneD-Viewer高亮显示某个或一组氨基酸，在另一个查看器（OneD-Viewer或Cn3D Viewer）中的相应部分就会自动高亮显示。

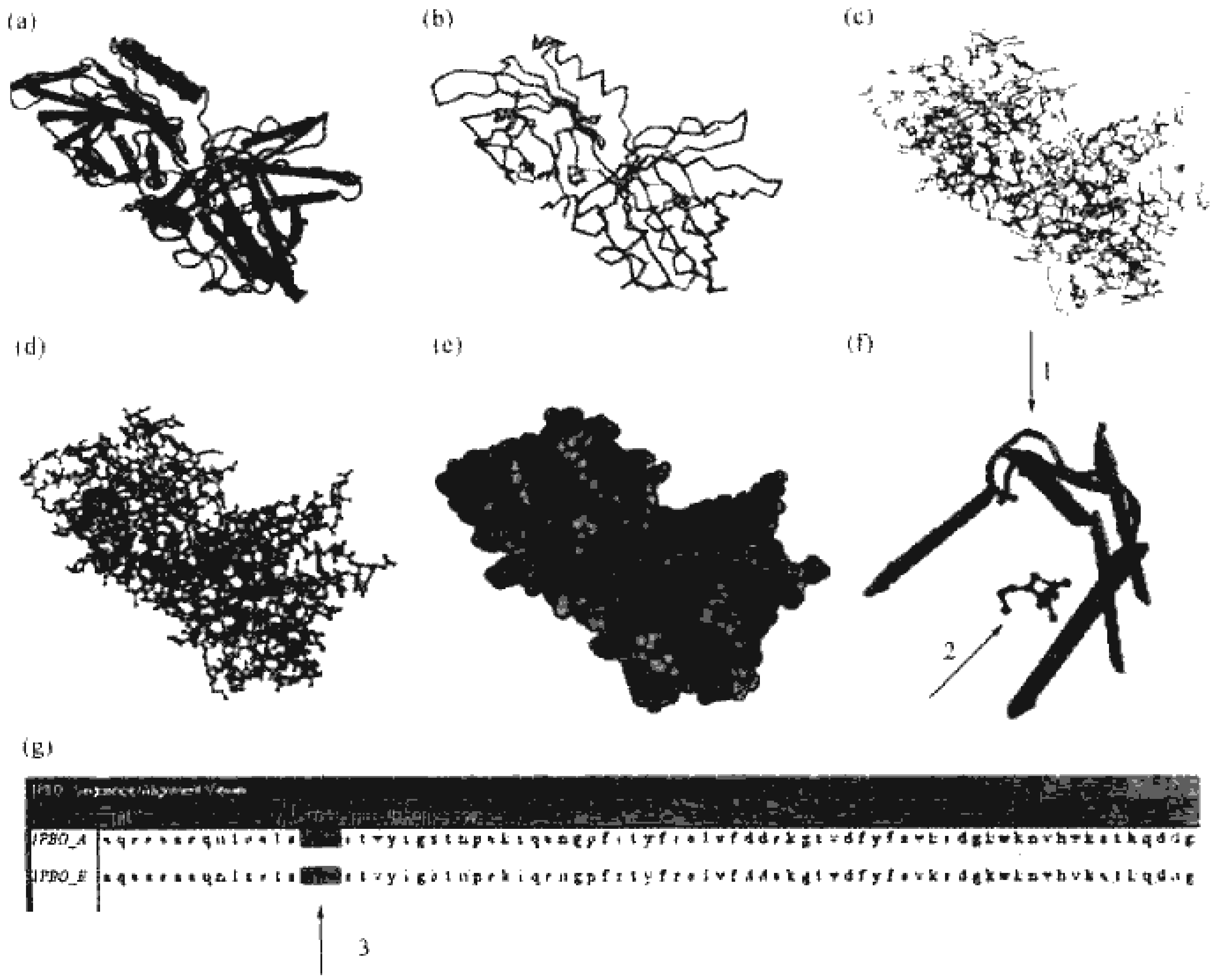


图9-16 NCBI的Cn3D Viewer显示的各种形式的增味剂结合蛋白（odorant-binding protein）二级结构——8束β折叠桶。(a) 螺纹状模型（worms）；(b) 管状模型（tubes）；(c) 线状模型（wire）；(d) 球框模型（ball and stick）；(e) 填充模型（space fill）。(f) 可以对结构进行旋转并放大任意区域；(g) 在Sequence/Alignment Viewer中显示蛋白质序列。点击含有保守模体GXW的序列（箭头3），蛋白质结构中的相应部分将会高亮显示（箭头1）。除此之外还有很多可视化选项，如显示结合位点的配体分子图像（箭头2）

除了可以观察单个蛋白质结构外，我们还可以对多个蛋白质的结构进行比对。在MMDB的"Structure Summary"页面中点击"Structure Neighbor"链接可以得到PDB中相关蛋白的列表（图9-17）。这个列表实际上是蛋白质矢量比对搜索工具（vector alignment search tool，VAST）的一部分。除增味剂结合蛋白（odorant-binding protein）外，再选择视黄醇结合蛋白（retinol-binding protein，1RLB）和β乳球蛋白（β-lactoglobulin，1BSO），然后在Cn3D Viewer中就可以看到这3个蛋白的结构了，同时相应的多序列比对会显示在二维查看器（DDV）中（图9-18）。VAST提供多种结构数据（框9-2）。

框 9-2 VAST 提供的信息

VAST 对发现的每个结构近邻，将提供下列信息。

① 选择框：通过选择框可以选取单个的结构近邻。

② PDB：结构近邻的 4 字符 PDB 标志码。

③ C：PDB 的链名。

④ D：MMDB 的结构域标志码。

⑤ RMSD：结构相似性度量，以 Å 为单位的方均根偏差。

⑥ NRES：两个结构中重叠的 α 碳原子的数量，这个量给出了对齐的长度。

⑦ %Id：序列比对中相同残基的百分含量。

⑧ 描述：来自 PDB 记录的一段文字。

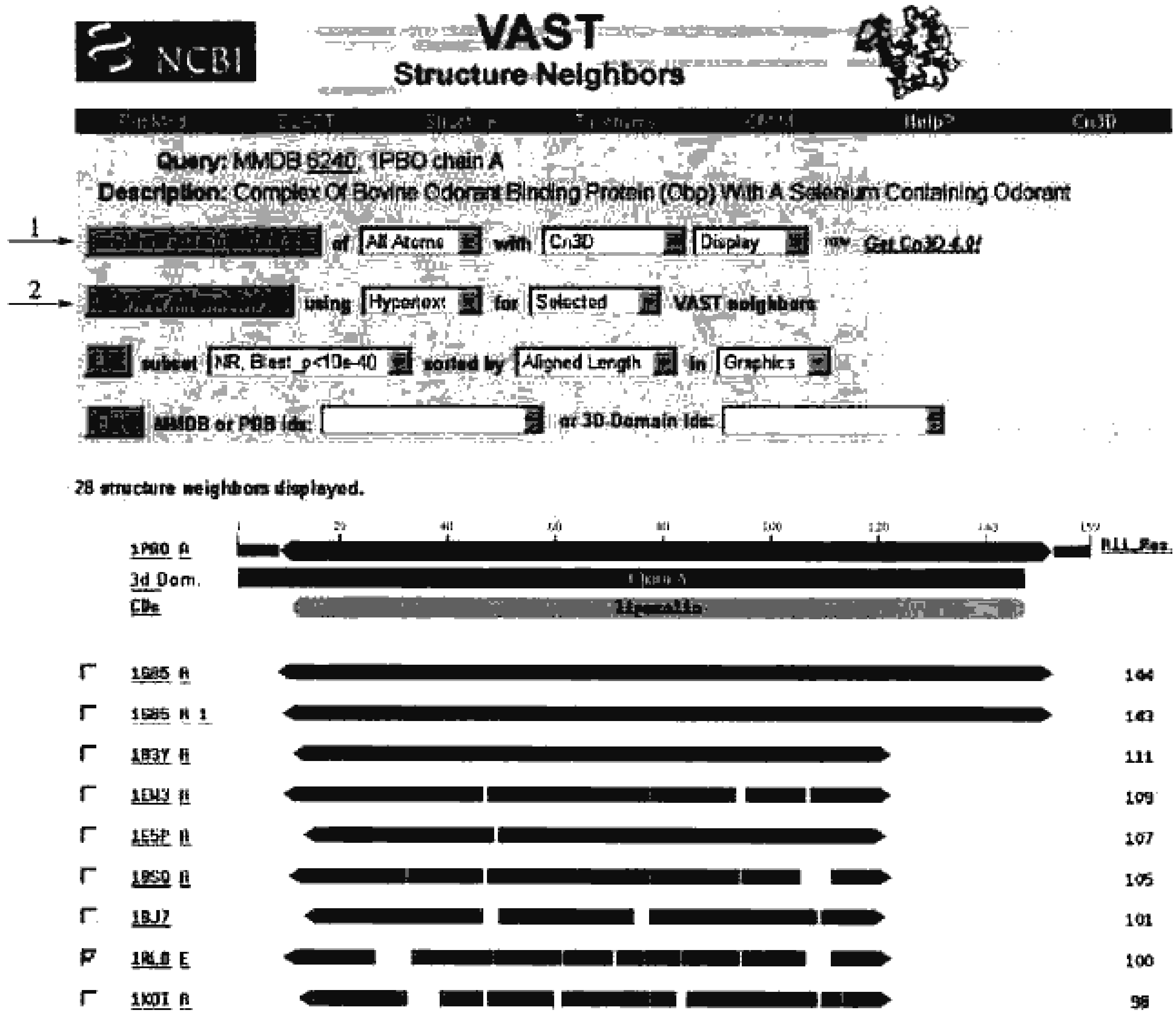


图 9-17 点击“Structure Neighbor”链接，可以进入相关蛋白的 VAST 页面。多个结构可以通过 Cn3D viewer 利用 VAST 工具进行比较（箭头 1），同时可以看到多序列比对结果（箭头 2）

另外还有很多基于网络的结构可视化工具。Swiss-PDB Viewer 提供了相当多的选项（包括 Ramachandran 图）将特定残基标记上颜色和一些建模程序。图 9-19 显示了以不同颜色标记的增味剂结合蛋白同源二聚体双链。这个结构是在含硒化合物配体存在的条件下用 X 射线衍射方法解析的。配体显示了每个亚基的“结合口袋”，这个结合位点被疏水残基包围，有利于结合疏水的气味分子。

有一些结构可视化程序利用一种网络浏览器插件 Chime，从而可以启用多种交互功能。图 9-20 和图 9-21 分别显示的是酵母 GAL4 蛋白质-DNA 复合体和肌红蛋白。同 Swiss-PDB Viewer 一样，这种软件提供了大量的查看选项。

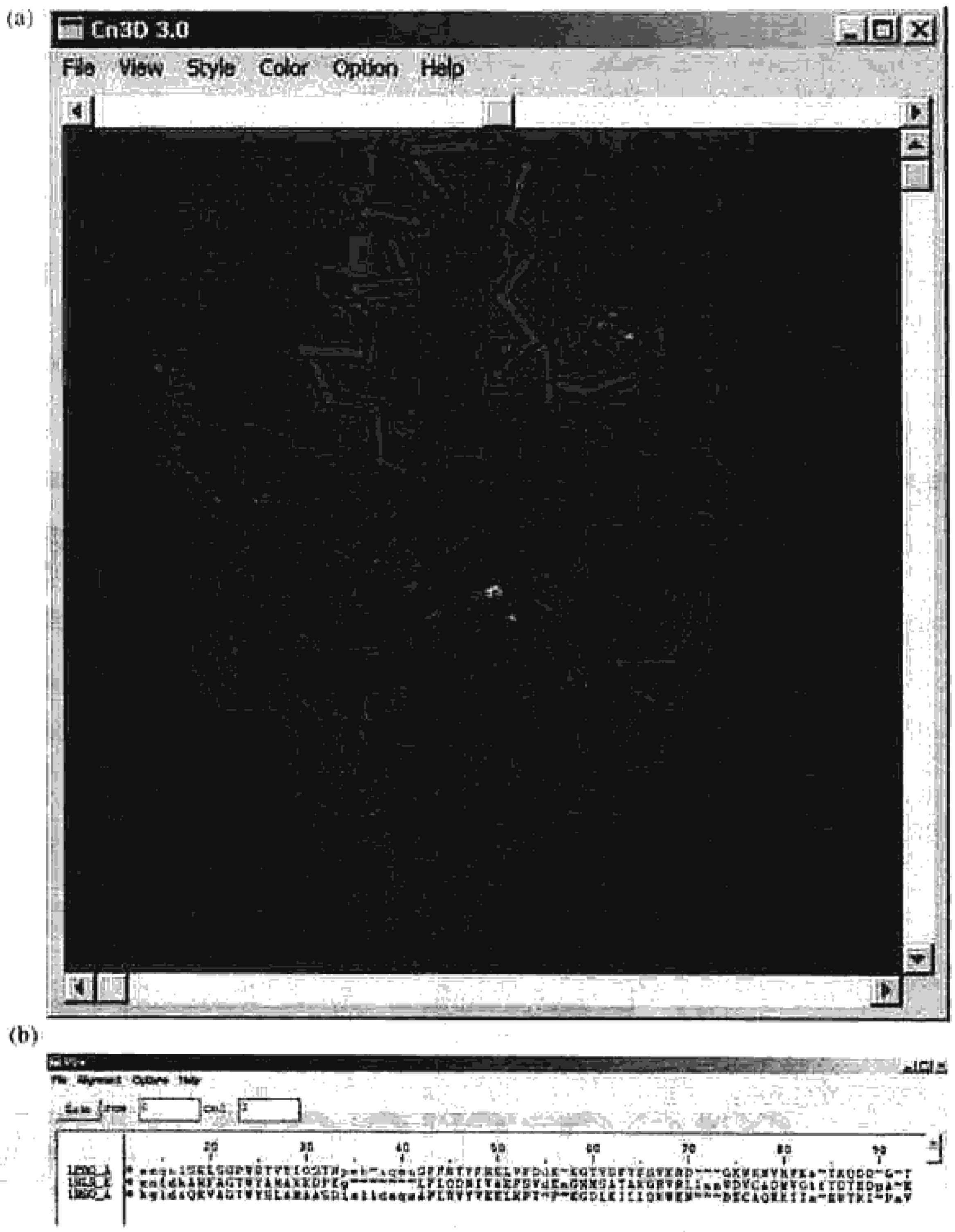


图 9-18 (a) 在 Cn3D 中比较从 VAST 选定的三种蛋白结构（OBP、RBP、β-lactoglobulin）；(b) 在 DDV 中显示的多序列比对。可以看到尽管这些蛋白质序列相似性很低，但是结构却极其相似

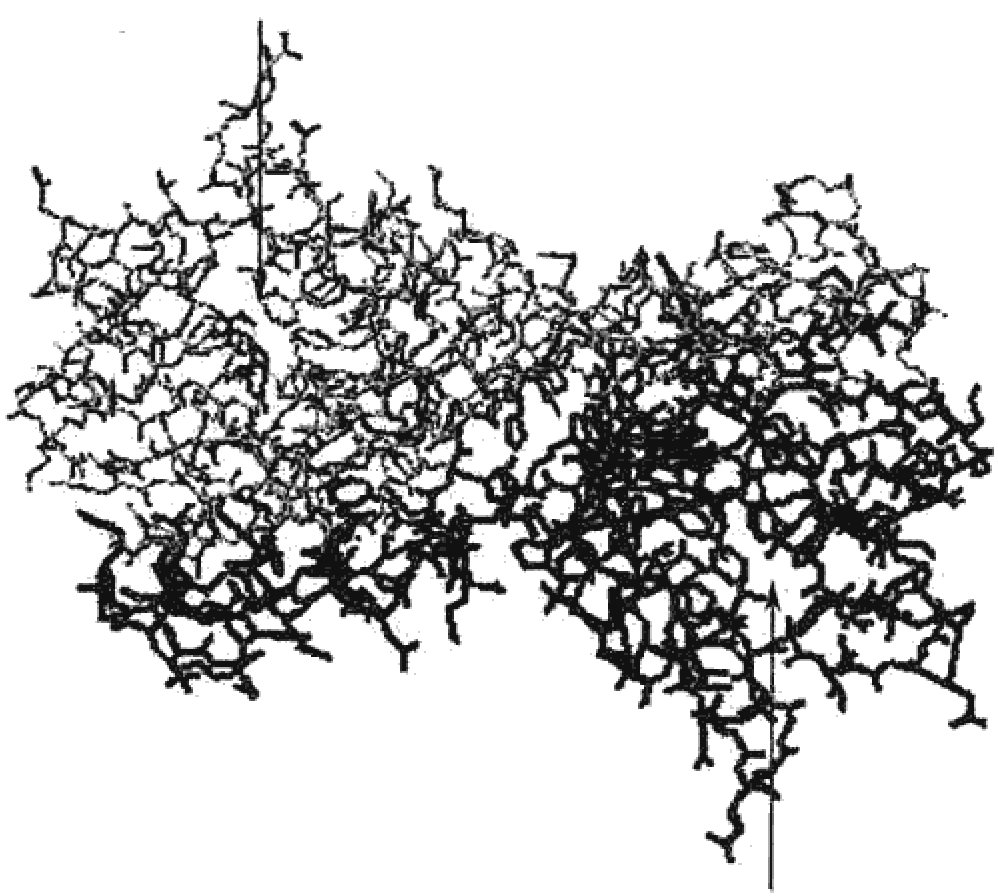


图 9-19 ExPASy 分析工具中包括了 Swiss-PDB Viewer（http://www.expasy.org/spdbv/）。与 Cn3D 相比，Swiss-PDB Viewer 提供了更多的功能用于复杂分析。这里 OBP 二聚体的两条链分别示以不同的颜色。这个结构是在含硒化合物配体存在的条件下用 X 射线衍射方法解析的，箭头显示了每个亚基的结合位点

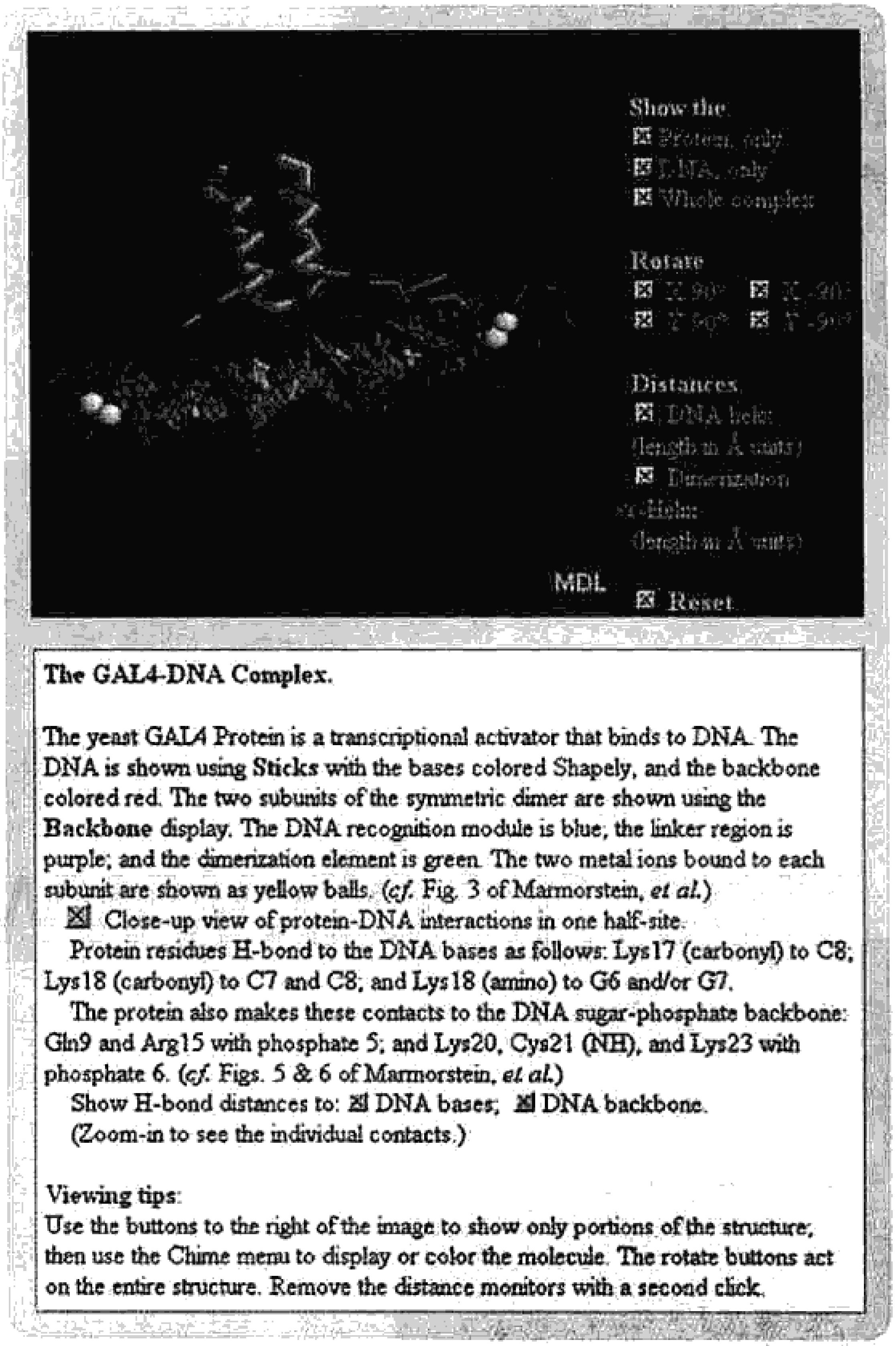


图 9-20 Chime 是一个网络浏览器插件，可以交互地显示 PDB 结构。这里显示的是酵母 GAL4 转录激活因子与 DNA 的结合。来自 http://info.bio.cmu.edu/course/03231/chime.tut/chime.html

(a)

(b)

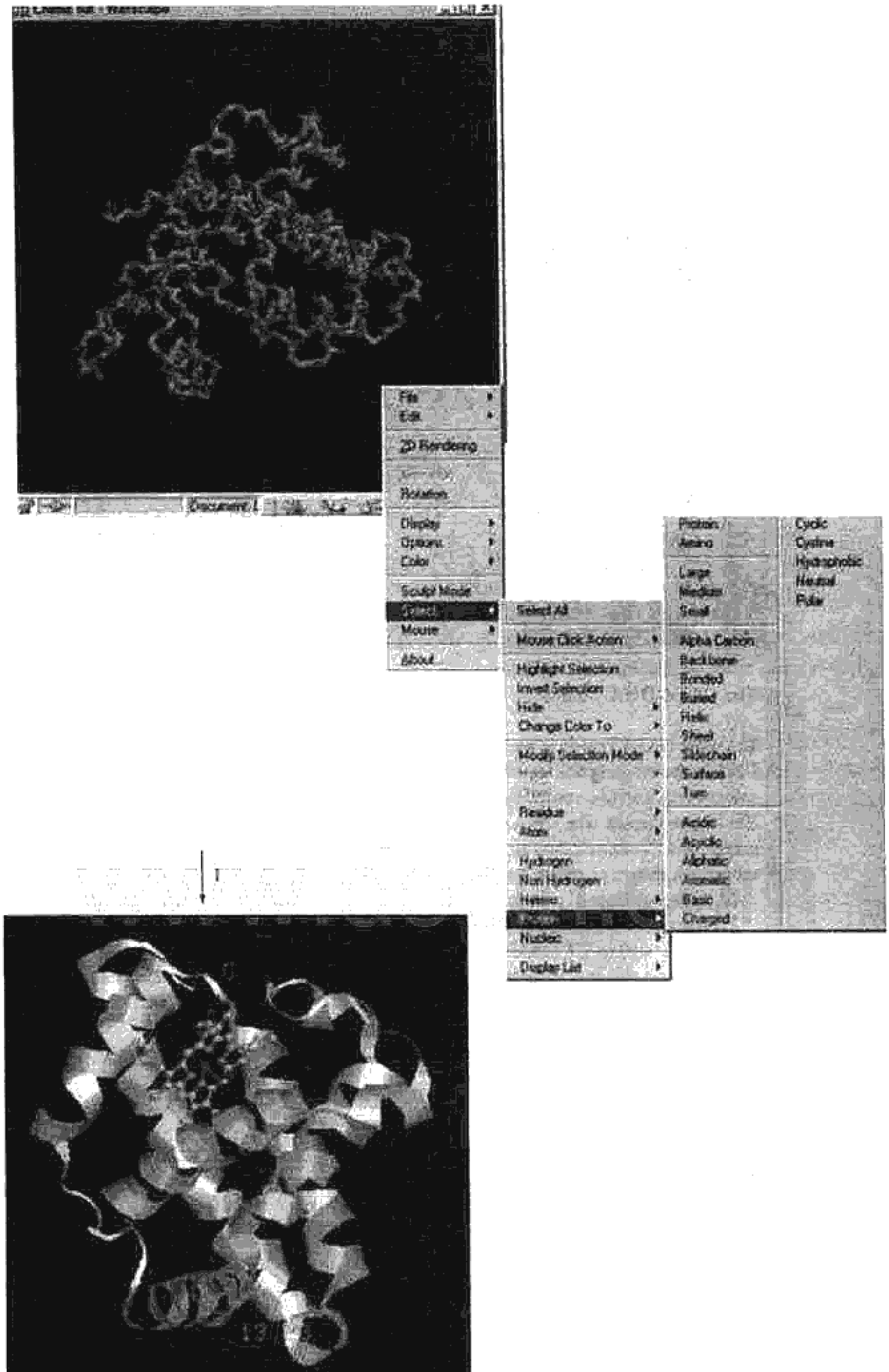


图 9-21 (a) 用 Chime 来观察肌红蛋白，下拉菜单中包括了许多查看选项。(b) 你可以看到血红素配体（箭头 1）或测量任意两个原子之间的距离（箭头 2）

9.2.2 蛋白折叠类型概况

现在我们已经学会了如何观察单个蛋白分子及对几个分子进行结构比较。另外有一些数据库是为研究蛋白质的折叠空间而建立的，即到底存在几种蛋白质折叠类型？有几种结构类型？为此我们将对 SCOP、CATH、Dali、FSSP 等几个数据库进行研究。

SCOP 数据库：http://scop.mrc-lmb.cam.ac.uk/scop/。

9.2.2.1 蛋白质分类系统：SCOP 数据库

基于一种分级分类系统（图 9-22）（Murzin 等，1995），蛋白质结构分类（SCOP）数据库提供了对蛋白质结构和进化关系的综合描述。位于分类层次顶部的是类（class），进一步划分为折叠子（fold）、超家族（super family）、家族（family）、蛋白质结构域（protein domain）、单个 PDB 蛋白结构记录。SCOP 数据库可以通过其分级结构导航图进行浏览，用关键字、PDB 标志码查询，或通过一个蛋白质序列进行同源搜索。

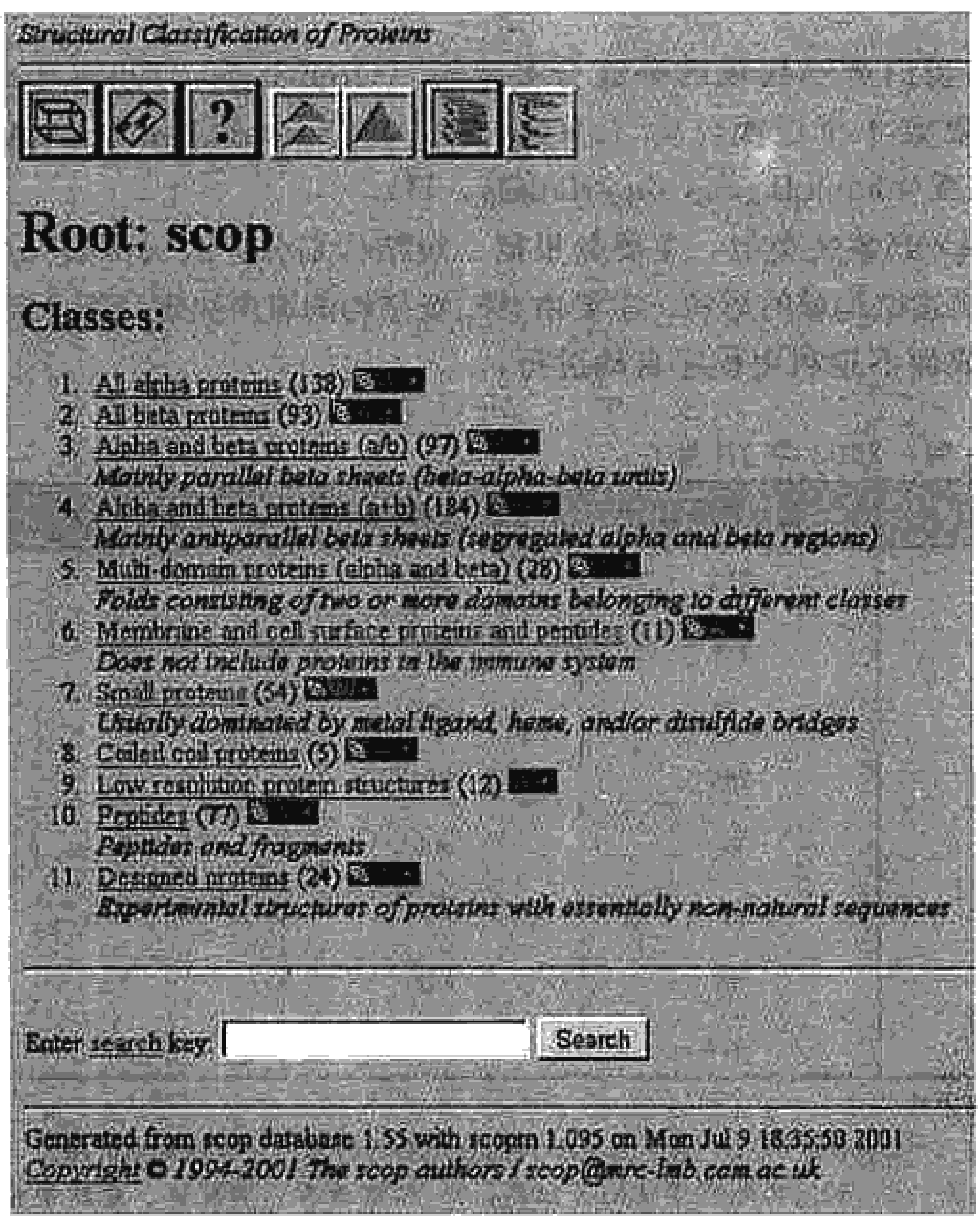


图 9-22 SCOP 数据库分类系统的结构

主要的类别都列举在表 9-4 中。结构层次中的折叠子描述的是具有相同的二级结构排列和拓扑取向的蛋白质类群。但是不同的蛋白质即使属于相同的折叠子在进化上也并不一定有联系。

折叠子以下是超家族层次。即使序列相似度很低，属于同一个超家族的蛋白质很可能在进化上有联系。下面是一个例子，SCOP 数据库中的载脂蛋白超家族既包括了我们前面提到的视黄醇结合蛋白（RBP）家族，还包括另一个以脂肪酸结合蛋白为代表的转运蛋白类群（FABPs）。与载脂蛋白一样，大部分 FABPs 是含量丰富、外分泌性的小分子（15ku），它

们与疏水性配体结合，通常在氨基端具有GXW模体。序列的两两比对找不出明显的匹配，但是FABP家族与RBP家族在进化上很可能是同源的。在SCOP数据库中，载脂蛋白超家族包含了3个类群：类RBP载脂蛋白（RBP-like lipocalins）、FABPs和凝血酶（thrombin）[有些研究者则认为calycin超家族由类RBP载脂蛋白、FABPs和抗生素蛋白avidins组成（Flower等，2000）]。

CATH位于http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html。
SSAP算法用于比较两个蛋白质的结构，可以从如下网页获得：http://www.biochem.ucl.ac.uk/cgi-bin/cath/GetSsapRasmol.pl。

在SCOP的分类层次中，属于同一个家族的蛋白质在进化上具有很明显的联系。通常这些蛋白质的结构相关，两两序列比对相似性高于30%。虽然在某些载脂蛋白家族和球蛋白家族内部成员序列相似性只有15%，但是进一步考虑结构和进化的关系后，这样的分类还是毫无疑义的。

9.2.2.2 CATH数据库

CATH是一个关于所有已知蛋白质结构域的分级分类系统（Orengo等，1997；Pearl等，2000）。CATH分类系统主要有4个层次：类（class，C），构架（architecture，A），拓扑（topology，T）和同源超家族（homologous superfamily，H）。

位于CATH最高层的类与SCOP系统类似，主要是根据二级结构划分：α、α+β、α/β、β等结构类型（表9-4）。构架层次描述结构域的形状，主要由其二级结构的取向决定（图9-23）。最终结构的判定是由专家人工完成而不是利用程序自动进行。

表9-4 SCOP数据库注释（1.59版，2002年10月）

类别	折叠子数目	超家族数目	家族数目	备注
全α蛋白	151	252	393	—
全β蛋白	110	205	337	—
α/β蛋白	113	185	438	1
α+β蛋白	208	295	454	2
多结构域蛋白	34	34	46	3
膜蛋白与细胞表面蛋白	12	19	31	4
小分子蛋白	58	83	128	5
总数	686	1073	1827	—

注：1. 每个折叠子含有一个到数十个超家族。

2. 备注栏中，1—主要是平行的β片层（β-α-β单位）；2—主要是反平行的β片层（分离的α区域和β区域）；3—含有多个结构域的折叠子可以属于不同的类；4—不包括免疫系统的蛋白质；5—经常含有金属配体、血红素和二硫键。

CATH的拓扑层次描述的是折叠家族的水平。家族的概念在SCOP和CATH中有所差异，SCOP家族的概念强调的是序列的相似性，而CATH的家族则强调二级结构的相似性。利用SSAP算法（Taylor和Orengo，1989）将蛋白质结构域聚类成不同的家族。属于同一个拓扑的蛋白质并不一定是同源的。与此对应，同源超家族层次则将那些具有同源性（如来自共同的祖先）的蛋白质聚类在一起。

对CATH数据库的搜索体现出这种分级分类系统的优越之处。在CATH首页可以用

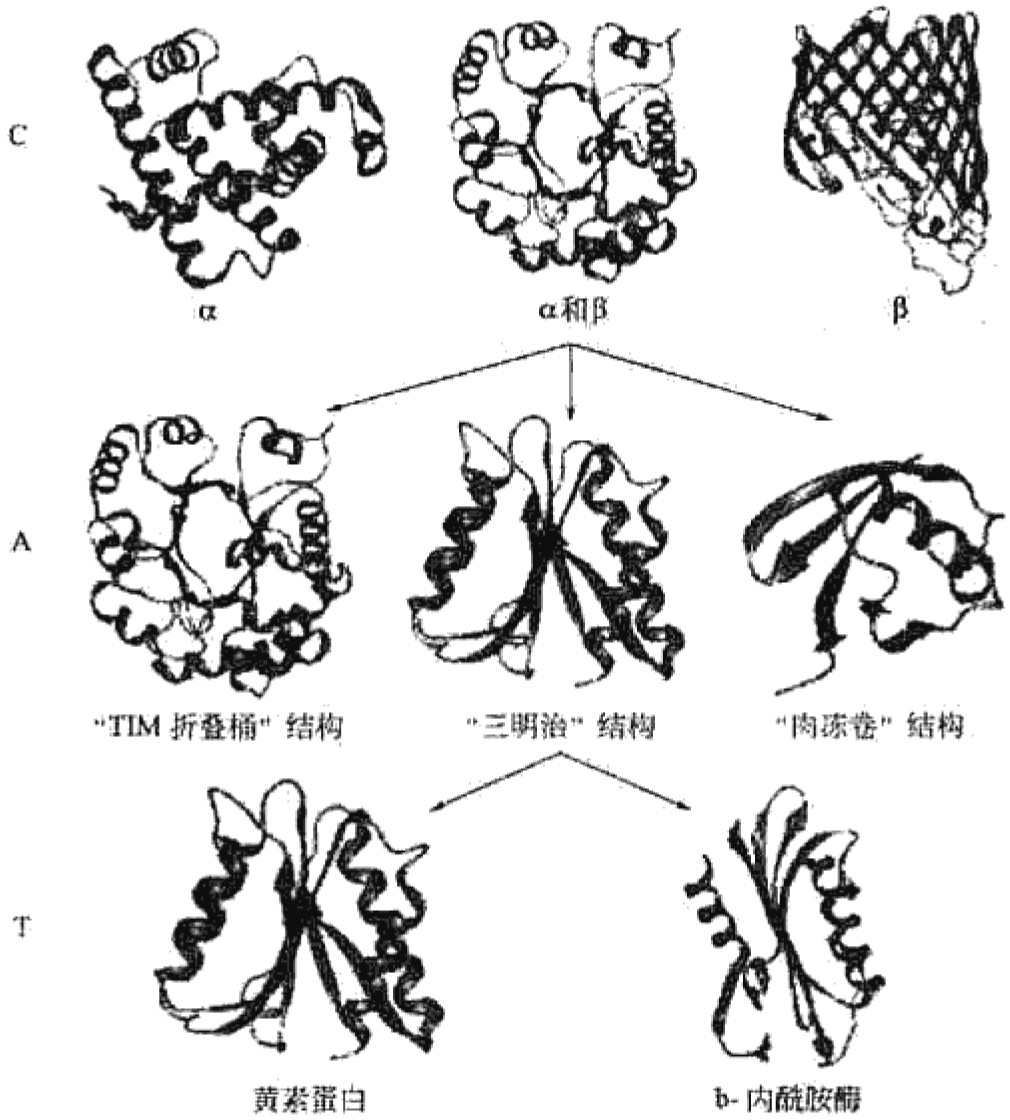


图 9-23　CATH 分类系统的结构层次：class、architecture、topology（fold family）和 homologous superfamily。来源于 http://www.biochem.ucl.ac.uk/bsm/cath_new/cath_info.html#C_Level

PDB 标志码、CATH 标志码或关键字进行搜索。如图 9-24 所示，结果对增味剂结合蛋白在 CATH 的各个分类层次上进行描述［图 9-25(a)］，同时结果还提供了关于其他载脂蛋白的链接［图 9-25(b)］。PDBsum 页面提供了几种结构查看器，指向 SwissProt、Pfam 数据库以及其他大量基于 PDB 的数据库链接（图 9-26 和图 9-27）。此外，我们还可以研究 CATH 的同源超家族字典（图 9-28）。总之 CATH 数据库提供了一套广而深的蛋白质结构方面的数据，并且将各种蛋白质结构置于分类学中的合适位置。

2000 年 9 月，Dali 分类系统已拥有超过 10000 个 PDB 记录，包含了 17100 条链，这些结构被划分为 5 个吸引子区域（attractor region），涵盖了超过 1300 个折叠类型，2500 个功能家族，3700 个结构域序列家族。Dali 位于 http://www2.ebi.ac.uk/dali/。

9.2.2.3　Dali 结构域字典

Dali 结构域字典的数字式分类系统对所有的 PDB 结构进行了分类（Dietmann 等，2000）。此分类系统的依据是一张对 PDB 中的所有蛋白质结构进行综合比较后构建的折叠空间图谱。建立了所有结构之间的相似性图表之后，按照不同的层次将这些结构进行聚类：

① 超二级结构模体（折叠空间中的吸引子）；

② 结构域的拓扑结构（折叠类型）；

③ 远缘同源（功能家族）；

④ 序列相似性大于 25%的同源（序列家族）。

Dali 中的每一个结构域都指定有一个分类编号，可以通过关键词进行查询，主查询页面见图 9-29。图 9-30 是关于增味剂结合蛋白（odorant-binding protein，DC_6_71_1_5）的部分记录。特别地，这个数据库整合了很多结构和序列方面的信息，如二级结构预测和溶剂可及性等。

FSSP 位于 http://www.ebi.ac.uk/dali/fssp。

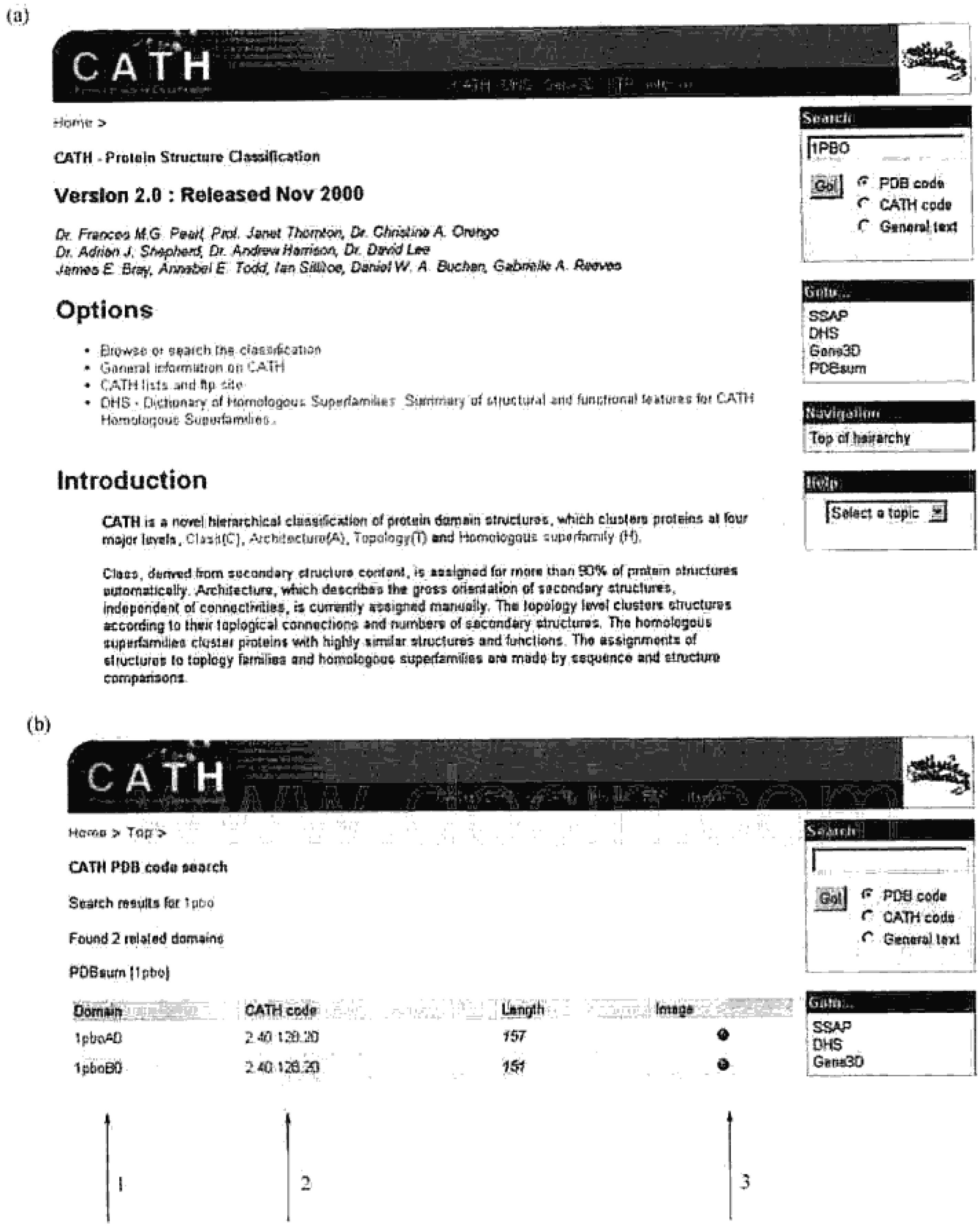
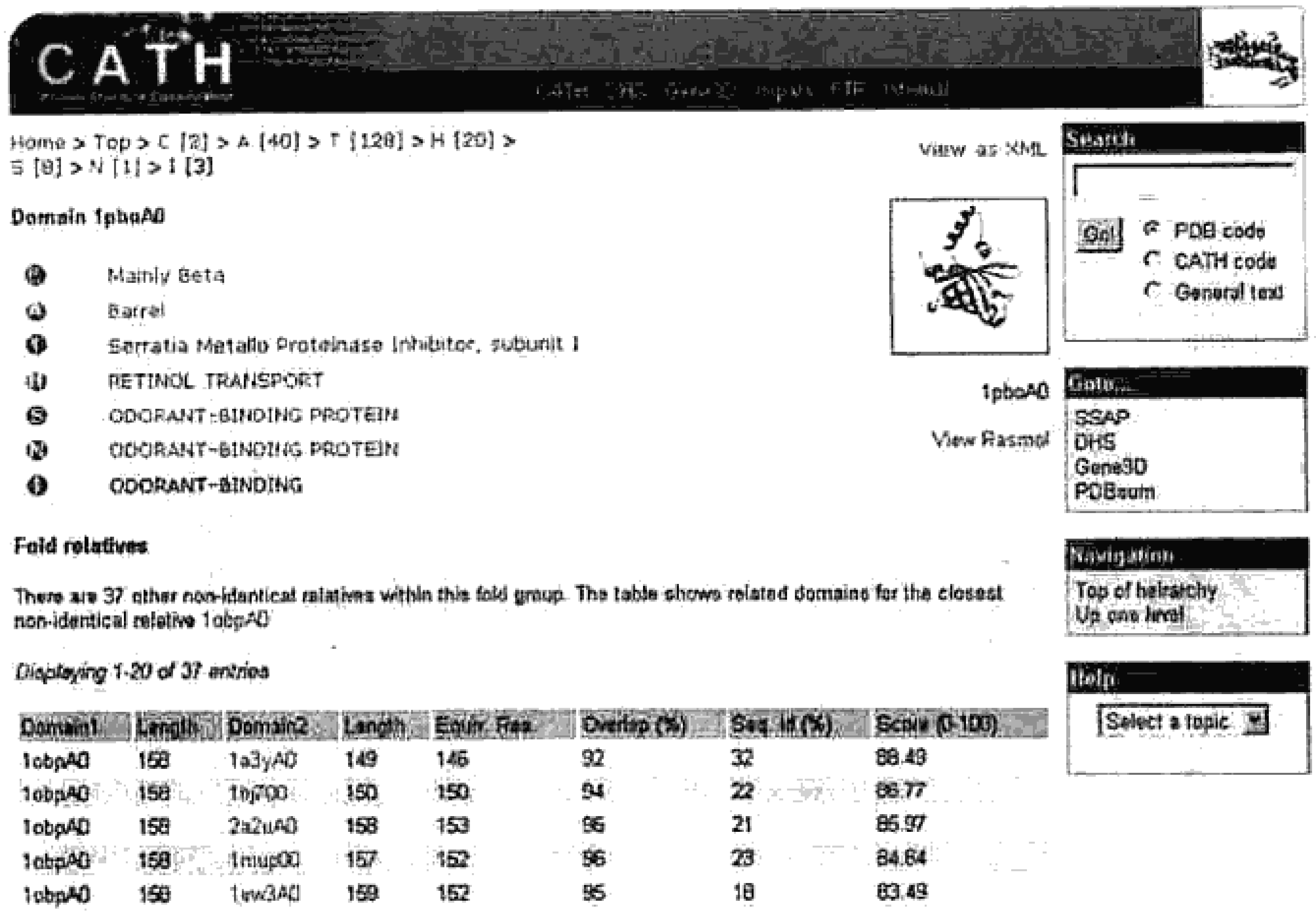


图 9-24 (a) 可以利用 PDB 标志码、CATH 标志码或文本在 CATH 首页对数据库进行搜索。(b) 用 PDB 标志码 1PBO 搜索的结果，其中包括了详细描述页面和静态图像的链接

(a)



(b)

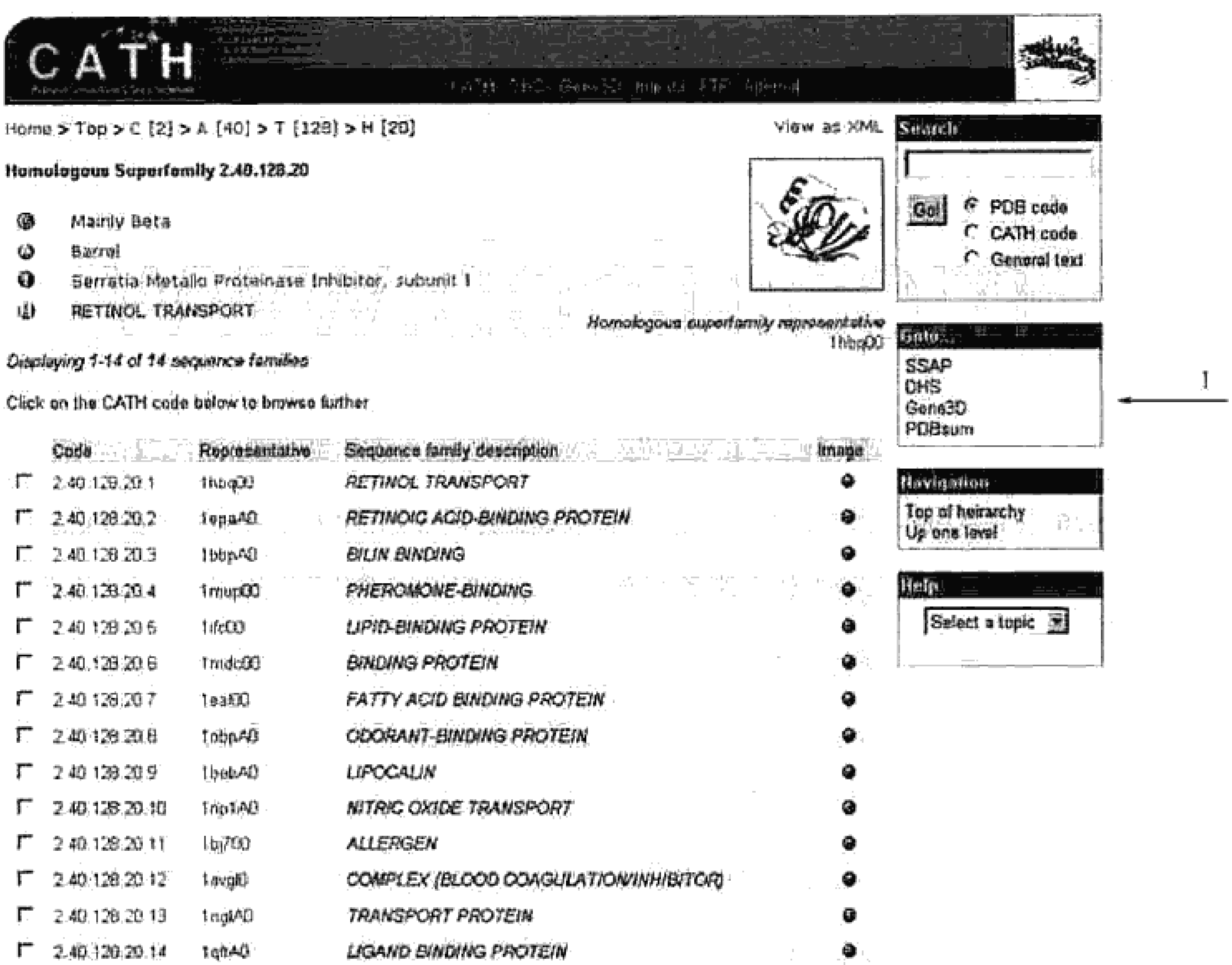


图 9-25　一次 CATH 数据库搜索（图 9-24）的结果 (a) 增味剂结合蛋白（odorant-binding protein）的一条链；(b) 邻近的载脂蛋白。结果还提供了其他一些链接，如 PDBSum（箭头 1）

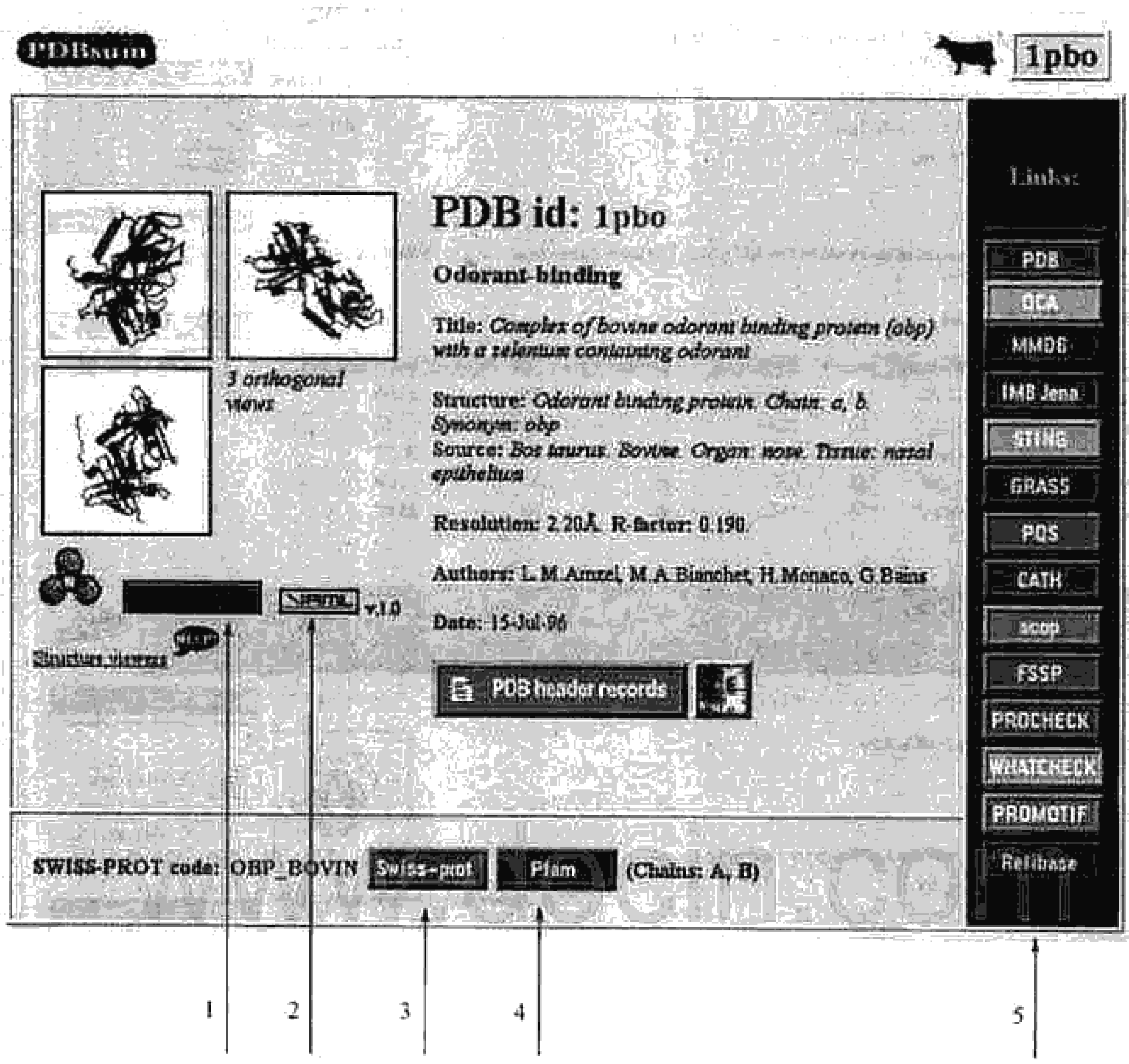


图 9-26 CATH 数据库中 Bovine OBP 的 PDBSu 记录提供了几种结构可视化工具（Rasmol，箭头 1，VRML，箭头 2），指向 SwissProt（箭头 3）、Pfam 数据库（箭头 4）以及其他大量基于 PDB 的数据库（箭头 5）链接。（更多的结果见图 9-27）

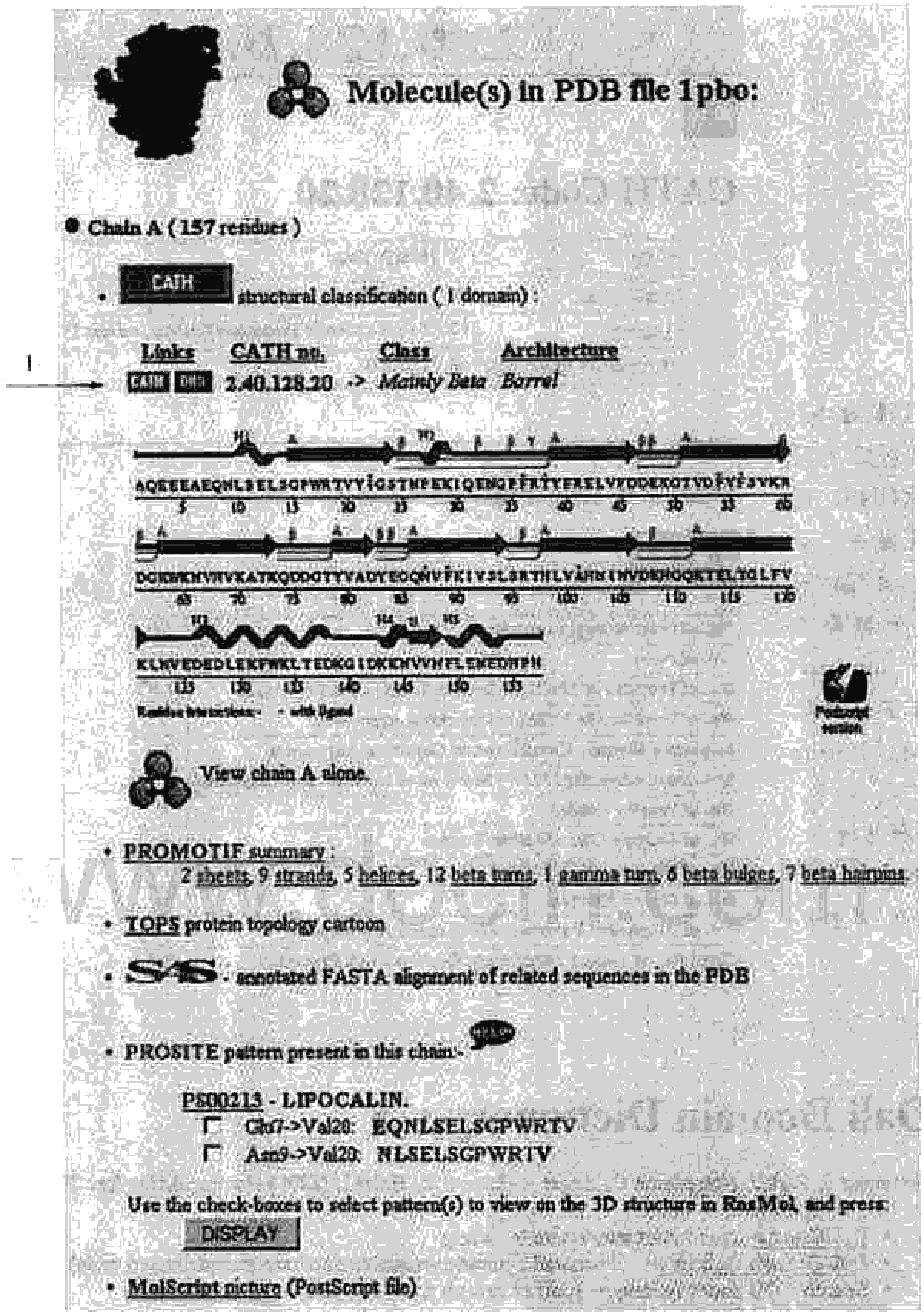


图 9-27 续图 9-26 CATH 数据库中 PDBsum 记录的输出页面，此结果显示了牛 OBP 的氨基酸序列和以卡通方式表示的 α 螺旋与 β 片层。另外，结果还提供了很多相关链接，包括 CATH 分类图式和同源超家族字典（箭头 1）

Dictionary of Homologous Superfamilies - 2.40.128.20 -

CATH Code: 2.40.128.20

Class	2	Mainly Beta
Architecture	40	Barrel
Topology	128	Serratia Metallo Proteinase Inhibitor, subunit I
Homologous superfamily	20	1euoA0

Summary Information

No. of N-level Representatives	34
No. of Pairs	561
No. of Proteins in Multiple Structural Alignment (corelist)	34 (4)
No. of Positions in Multiple Structural Alignment	311
Sequence Identity Cutoff used to Generate Alignment	25
Structural Similarity (SSAP) Score used to Generate Alignment	80
No. of Prosite Entries	33
No. of Enzyme Classification Entries	0
No. of SWISS-PROT Entries	34
No. of Ligand Entries	26
No. of Alignment Positions with Ligand Interactions	77
Total No. of Ligand Interacting Residues (All Proteins)	238

图 9-28 CATH的bovine OBP记录包括了同源超家族字典（DHS）链接。在DHS页面中，除了CATH分类编码（CATH code）和概要信息（summary information）外，还有蛋白质结构域的多重比对、Prosite类型列表、E.C编号、SwissProt描述和关键字、蛋白质-配体结合类型、位置特异性配体信息矩阵等

Dali Domain Dictionary v.3

Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L (2001) Nucleic Acids Res 29, 55-57.

- Dali Classification - start interactive browsing
- Dali Domain Definitions - download computer-readable database (tab-delimited table)
- Search PDB codes (example: 1ppt) or protein names (example: tyrosine kinase)

lipocalin [submit] [reset]

图 9-29 Dali结构域字典（http://www.ebi.ac.uk/dali/domain）是一个数字化的蛋白质结构分类系统，根据结构、功能、序列的相似性（Dietmann等，2001）对蛋白质进行分类。此系统的基础是一张对PDB中的所有蛋白质结构进行综合比较后构建的折叠空间图谱

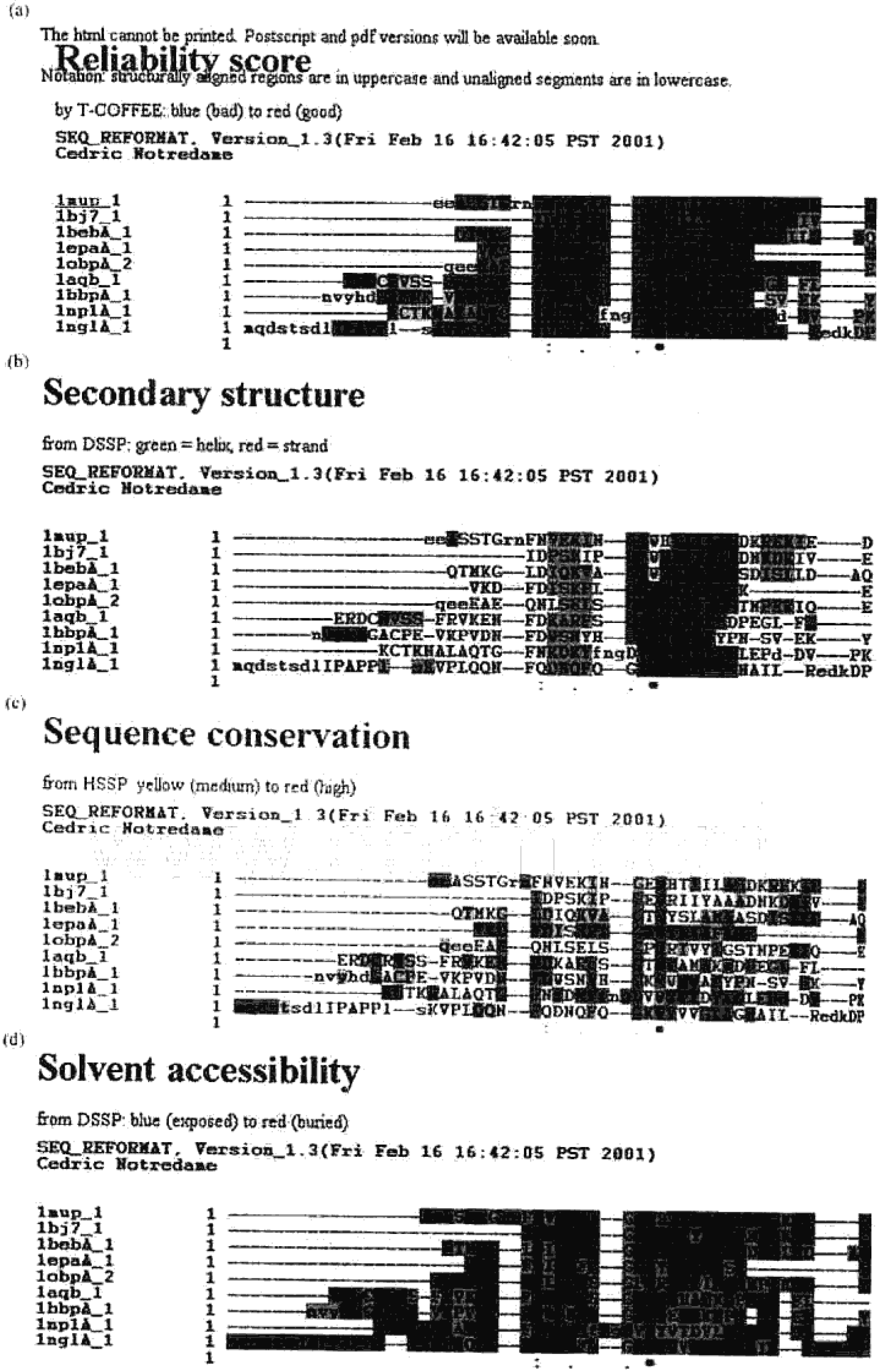
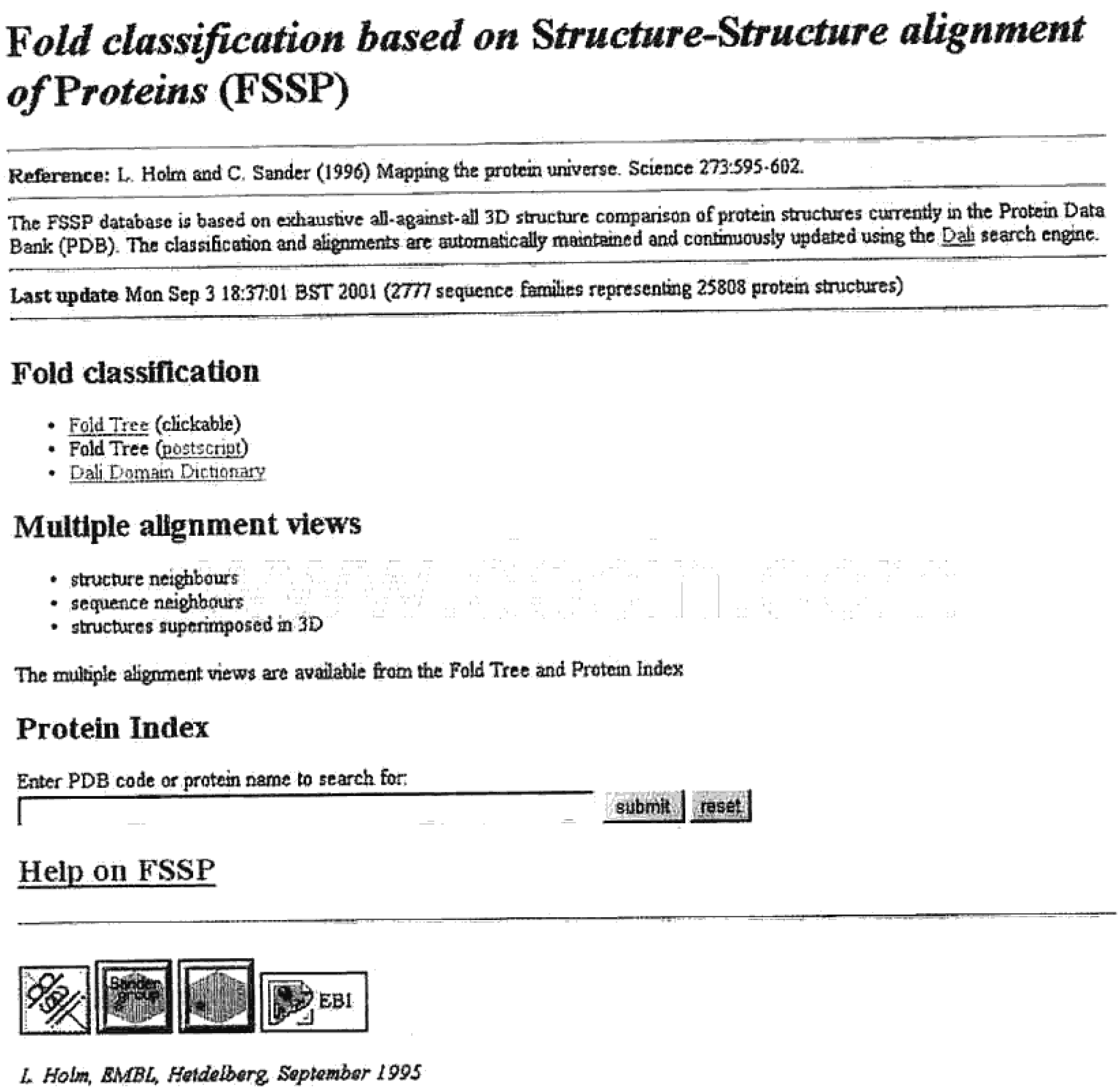


图 9-30 Dali 服务器提供的基于结构信息的多序列比对信息，包括（a）可信度、（b）二级结构、（c）序列保守性、（d）溶剂可及性

9.2.2.4 FSSP数据库

FSSP的全称是基于蛋白质结构比对的折叠分类（the fold classification based on structure-structure alignment of proteins）。此数据库将PDB中的所有结构进行了综合的比较（图9-31）（Holm和Sander，1996a），保存其中所有长度大于30个残基的结构。这些结构被划分为一系列“代表集合”，每一个代表集合内结构的序列相似性不大于25%；然后根据折叠类型树形表（图9-32）对每一个代表集合内的折叠进行分类。点击增味剂结合蛋白（odorant-binding protein）的记录，可以得到一组结构近邻（图9-33），选择其中的一部分后，还可以进行多序列比对（图9-34）并观察它们三维结构的重合部分。



Fold classification based on Structure-Structure alignment of Proteins (FSSP)

Reference: L. Holm and C. Sander (1996) Mapping the protein universe. Science 273:595-602.

The FSSP database is based on exhaustive all-against-all 3D structure comparison of protein structures currently in the Protein Data Bank (PDB). The classification and alignments are automatically maintained and continuously updated using the Dali search engine.

Last update Mon Sep 3 18:37:01 BST 2001 (2777 sequence families representing 25808 protein structures)

Fold classification

- Fold Tree (clickable)
- Fold Tree (postscript)
- Dali Domain Dictionary

Multiple alignment views

- structure neighbours
- sequence neighbours
- structures superimposed in 3D

The multiple alignment views are available from the Fold Tree and Protein Index

Protein Index

Enter PDB code or protein name to search for:

submit reset

Help on FSSP

EBI

L. Holm, EMBL, Heidelberg, September 1995

图9-31 FSSP数据库属于欧洲生物信息学研究所（http://www.ebi.ac.uk:80/dali/fssp/）（Holm和Sander，1996）

255.1.1.1.1.1	1f3uA	alignment	transcription initiation factor iif, beta subunit fragment
255.1.2.1.1.1	1f3uB	alignment	transcription initiation factor iif, beta subunit fragment
256.1.1.1.1.1	1yua	alignment	topoisomerase i fragment
257.1.1.1.1.1	2omf	alignment	matrix porin outer membrane protein f (matrix porin, ompf p
257.1.1.1.2.1	1e54A	alignment	outer membrane porin protein 32 (omp32) omp32
257.1.1.1.3.1	3prn	alignment	porin Mutant biological_unit
257.1.1.1.4.1	2por	alignment	Porin (crystal form b)
257.1.1.2.1.1	2mprA	alignment	maltoporin fragment (lam-b, mal-1) biological_unit
257.1.1.2.1.2	1a0tP	alignment	sucrose-specific porin biological_unit
257.1.1.3.1.1	1fepA	alignment	ferric enterobactin receptor (fepa)
257.1.1.3.2.1	1by5A	alignment	ferric hydroxamate uptake protein (fhua) ferrichrome
257.1.2.1.1.1	1qd5A	alignment	outer membrane phospholipase a Mutant
257.1.3.1.1.1	4bcl	alignment	bacteriochlorophyll a protein biological_unit
257.1.4.1.1.1	1qj8A	alignment	outer membrane protein x Mutant
257.1.4.2.1.1	1bxwA	alignment	outer membrane protein a fragment Mutant
257.1.5.1.1.1	1swuA	alignment	streptavidin Mutant
257.2.1.1.1.1	1dwnA	alignment	phage coat protein
257.2.1.2.1.1	1gbeA	alignment	bacteriophage q beta capsid
257.2.1.3.1.1	1msc	alignment	Bacteriophage ms2 unassembled coat protein dimer (translati
258.1.1.1.1.1	1bj7	alignment	d 2 (dander major allergen bda20, dermal allergen bda20) bi
258.1.1.1.2.1	1mup	alignment	Major urinary protein complex with 2-(sec-butyl) thiazoline
258.1.1.1.3.1	1bebA	alignment	beta-lactoglobulin biological_unit
258.1.1.1.4.1	1epaA	alignment	Epididymal retinoic acid-binding protein (androgen dependen
258.1.1.2.1.1	1bbpA	alignment	Bilin binding protein (BBP)
258.1.1.2.2.1	1aqb	alignment	retinol-binding protein (rbp)
258.1.1.3.1.1	1np1A	alignment	nitrophorin 1 (np1) biological_unit
258.1.1.4.1.1	1nglA	alignment	ngal fragment (hngal, hnl) biological_unit
258.1.1.5.1.1	1qftA	alignment	female-specific histamine binding protein 2 female-specific ← 1
258.1.1.6.1.1	1obpA	alignment	odorant-binding protein (obp)
258.1.2.1.1.1	1hmt	alignment	Fatty acid binding protein (human muscle, m-fabp) complexed
258.1.3.1.1.1	1avgI	alignment	thrombin triabin
258.2.1.1.1.1	1smpI	alignment	serratia metallo proteinase erwinia chrysanthemi inhibitor
259.1.1.1.1.1	1bvqA	alignment	4-hydroxybenzoyl coa thioesterase
259.1.1.2.1.1	1c8uA	alignment	acyl-coa thioesterase ii
259.1.1.3.1.1	1mkaA	alignment	beta-hydroxydecanoyl thiol ester dehydrase (beta-hydroxydec
260.1.1.1.1.1	1ffkU	alignment	23s rrna 5s rrna ribosomal protein l2 (50s ribosomal protei
261.1.1.1.1.1	1bv1	alignment	bet v 1 (major pollen allergen bet v 1-a)
261.1.1.2.1.1	1em2A	alignment	mln64 protein fragment Mutant
261.1.1.3.1.1	1fvzA	alignment	phosphatidylinositol transfer protein (pitp)
261.1.2.1.1.1	3pmgA	alignment	alpha-d-glucose-1,6-bisphosphate (phosphoglucomutase)
261.2.1.1.1.1	1jenA	alignment	s-adenosylmethionine decarboxylase (beta chain) (adometdc,
261.3.1.1.1.1	1d8iA	alignment	mrna triphosphatase cet1
261.4.1.1.1.1	1eaf	alignment	Dihydrolipoyl transacetylase (catalytic domain (residues 38
261.4.1.1.2.1	3cla	alignment	Type III chloramphenicol acetyltransferase (CAT-III*) compl
262.1.1.1.1.1	1fmaE	alignment	molybdopterin convertin factor, subunit 1 molybdopterin con
262.2.1.1.1.1	2fua	alignment	l-fuculose-1-phosphate aldolase
263.1.1.1.1.1	1g8jB	alignment	arsenite oxidase fragment arsenite oxidase fragment
263.1.1.2.1.1	1fqtA	alignment	rieske-type ferredoxin of biphenyl dioxygenase (biphenyl di
263.1.1.3.1.1	1rie	alignment	rieske iron-sulfur protein fragment (ubiquinol-cytochrome c
263.1.1.4.1.1	1rfs	alignment	rieske protein fragment (rieske iron-sulfur protein, risp)

图 9-32 FSSP 的折叠树实际上是标上标志码的折叠子列表，可以通过名字或标志码进行查询，图中显示了 OBP 的位置（箭头 1）

FSSP: select structural neighbours of 1obpA

Please cite: L. Holm and C. Sander (1996) Science 273(5275):595-60.

Select (check) structural neighbours to display

[3D superimposition] [Multiple alignment (wide)] [Multiple alignment (narrow)] [Multiple families (wide)]
[Multiple families (narrow)] [Reset selection]

	STRID2	Z	RMSD	LALI	LSEQ2	%IDE	PROTEIN
☐	1obpA	27.2	0.0	156	159	100	odorant-binding protein (obp)
☐	1pboB	23.0	0.3	145	151	99	odorant binding protein (obp)
☐	1obpB	22.0	0.8	145	155	100	odorant-binding protein (obp)
☐	1pboA	21.7	0.8	145	157	99	odorant binding protein (obp)
☐	1dzpB	16.7	1.8	115	147	42	odorant-binding protein (pig obp)
☐	1a3yB	16.7	1.8	115	147	42	odorant binding protein biological_unit
☐	1dzmB	16.5	1.8	115	147	42	odorant-binding protein (pig obp)
☐	1dzjA	16.5	1.9	115	148	42	odorant-binding protein (pig obp)
☐	1e00A	16.4	1.9	115	149	42	odorant-binding protein synonym pig obp
☐	1a3yA	16.4	1.8	115	149	42	odorant binding protein biological_unit
☐	1dzpA	16.4	1.9	115	149	42	odorant-binding protein (pig obp)
☐	1dzmA	16.3	1.9	115	149	42	odorant-binding protein (pig obp)
☐	1e00A	16.3	1.9	115	149	42	odorant-binding protein synonym pig obp
☐	1dzjB	15.3	1.8	115	148	42	odorant-binding protein (pig obp)
☐	1dzkB	15.3	1.8	115	148	42	odorant-binding protein (pig obp)
☐	1e00B	15.3	1.8	115	147	42	odorant-binding protein synonym pig obp
☐	1e00B	15.2	1.8	115	148	42	odorant-binding protein synonym pig obp
☐	1dzkA	15.1	1.9	115	148	42	odorant-binding protein (pig obp)
☐	1e02B	15.1	2.0	116	147	41	odorant-binding protein synonym pig obp
☐	1hqpA	15.0	1.8	115	149	42	odorant-binding protein
☐	1bj7	15.0	2.2	117	150	26	d 2 (dander major allergen bda20, dermal allergen bda20
☐	1e02A	14.9	1.9	115	149	42	odorant-binding protein synonym pig obp
☐	1mup	13.8	2.4	118	157	28	Major urinary protein complex with 2-(sec-butyl) thiazo
☐	1bebA	12.0	2.3	116	156	14	beta-lactoglobulin biological_unit
☐	1epbA	8.7	2.8	104	151	12	Epididymal retinoic acid-binding protein (androgen depe
☐	1aqb	7.9	3.0	109	175	14	retinol-binding protein (rbp)
☐	1np1A	7.4	3.3	111	184	8	nitrophorin 1 (np1) biological_unit
☐	1nglA	7.4	4.1	110	179	11	ngal fragment (hngal, hnl) biological_unit
☐	1bbpA	6.7	3.8	111	173	13	Bilin binding protein (BBP)
☐	1qftA	5.8	3.4	107	175	10	female-specific histamine binding protein 2 female-spec
☐	1bxwA	5.4	3.7	101	172	3	outer membrane protein a fragment Mutant

图 9-33 使用 FSSP 工具可以选出一个蛋白的结构近邻供深入分析

FSSP: structural neighbours of 1obpA

Please cite: L. Holm and C. Sander (1996) Science 273(5275):595-60.

Structural alignment by Dali

Notation: Uppercase: structurally equivalent with 1obpA; lowercase: structurally non-equivalent with 1obpA.

```
1obpA    QEEE........AEQNLSELSGPWRTVYIGSTNPEKIQE.NGPFRT.YFRELVFDDEKGT
1obpA    QEEE........AEQNLSELSGPWRTVYIGSTNPEKIQE.NGPFRT.YFRELVFDDEKGT
1bebA    .........qtmKGLDIQKVAGTWYSLAMAASDISLLDAqSAPLRV.YVEELKPT.PEGD
1aqb     erdcrvssfrvkENFDKARFSGTWYAMAKKDPEG.......LFLQdnIVAEFSVD.ENGH

1obpA    VDFYFSVKRD.....GKWKNVHVKATKQDDG.TYVADY........EGQNVFKIVSLS.R
1obpA    VDFYFSVKRD.....GKWKNVHVKATKQDDG.TYVADY........EGQNVFKIVSLS.R
1bebA    LEILLQKWEN.....GECAQKKIIAEKTKIPaVFKIDA........LNENKVLVLDTDyK
1aqb     MSATAKGRVRllnnwDVCADMVGTFTDTEDPaKFKMKYwgvasflqKGNDDHWIIDTDyD

1obpA    THLVAHNINVDKHGQ..TTELTGLFVKLNVEDEDLEKFWKLTEDKGIDKKNVVNFLENED
1obpA    THLVAHNINVDKHGQ..TTELTGLFVKLNVEDEDLEKFWKLTEDKGIDKKNVVNFLENED
1bebA    KYLLFCMENS.aepE..QSLVCQCLVRT.pevddealekfdkalkalpmhirlsfnptql
1aqb     TYAVQYSCRLqnldgtcADSYSFVFARD..phgfspevqkivrqrqeelclarqyrlith

1obpA    HPHPE
1obpA    HPHPE
1bebA    EEQC.
1aqb     NGYCD
```

return to FSSP home page / Dali Domain Dictionary

(C) L. Holm, EMBL-EBI, Hinxton, May 1996

图 9-34 FSSP 中 OBP（1obpA）与其结构近邻（β-乳球蛋白和视黄醇结合蛋白）的多重比对

9.3 用于结构研究的计算生物学方法

9.3.1 同源（比较）模建

SwissProt/TrEmBL 大约收录了 100 万个蛋白质序列（2003 年 6 月），而 PDB 所收录的蛋白质结构却只有 2 万个。

对于大多数蛋白质，其结构模型的确定依赖于实验以外的计算生物学方法。随着越来越多的蛋白质结构被 X 射线晶体学方法和核磁共振方法解析，基于已有结构知识的建模与评估方法的可信度也得到了提高（Jones，2001；Baker 和 Sali，2001）。同源模建是结构基因组学领域的一个非常基本的方法。

同源模建主要包括以下 4 步（Marti-Renom 等，2000）。

① 首先指定折叠类型并选择模板。利用 BLAST 和 PSI-BLAST 之类工具搜索同源蛋白质序列与结构。可以在前面提到的一些数据库中（PDB，CATH，SCOP 等）对目标进行搜索，根据结果分析从而确定结构保守区与可变区。通常情况下，结构可变区对应位于蛋白质外部的卷曲和转角。

② 将目标与模板对齐。对于远缘的蛋白质，很难得到精确的比对结果。如果目标与模板有 30%以上的序列相似性且匹配区域长度足够（>60 个残基），那么它们很可能具有相似的结构。第 10 章提到的多序列比对方法将显得非常有用。

③ 建立模型。可以使用多种方法，如刚体装配和片段匹配。

④ 评估模型。

在进行同源模建时主要可能会发生下列错误。

① 错误的侧链装配。

② 在正确对齐的区域中发生扭曲。

③ 在目标未匹配区域发生错误。

④ 序列比对时出错。

⑤ 使用了错误的模板。

预测蛋白质结构的准确程度和目标、模板序列的相似度有很大关系（图 9-36）。当两个序列的相似度超过 50%时，模型的质量将非常好。在这种情况下，主链原子坐标的均方根偏差可以低于 0.1nm。当序列相似度介于 30%～50%时，模型准确度开始下降，序列相似度低于 30%时，模型的错误会急剧上升。从头预测的方法只能得到低分辨率的结构模型。

有很多同源模建的网络服务器，像 ExPASy 的 SWISS-MODEL（图 9-37），Columbia 的 Predict Protein 服务器（图 9-38 和图 9-39），CMBI 的 WHAT IF 和 San Diego 的 Protein Structure Homology Modeling Server。其他一些资源列于表 9-6。

Bystroff 和 Shao（2002）将 Rosetta Stone 和相关的一些从头预测方法整合，建立了一个蛋白质结构预测服务器：http://www.bioinfo.rpi.edu/~bystrc/hmmstr/server.html。

9.3.2 从头预测

在没有明显的同源时，我们可以考虑使用从头预测的方法来获得蛋白质结构模型（图 9-35）。由自由能全局最小化判据对蛋白质进行模拟折叠，并不与已知的蛋白质结构进行比较（Osguthorpe，2000；Simons 等，2001）。尽管分辨率很低，这仍不失为一种得到结构模

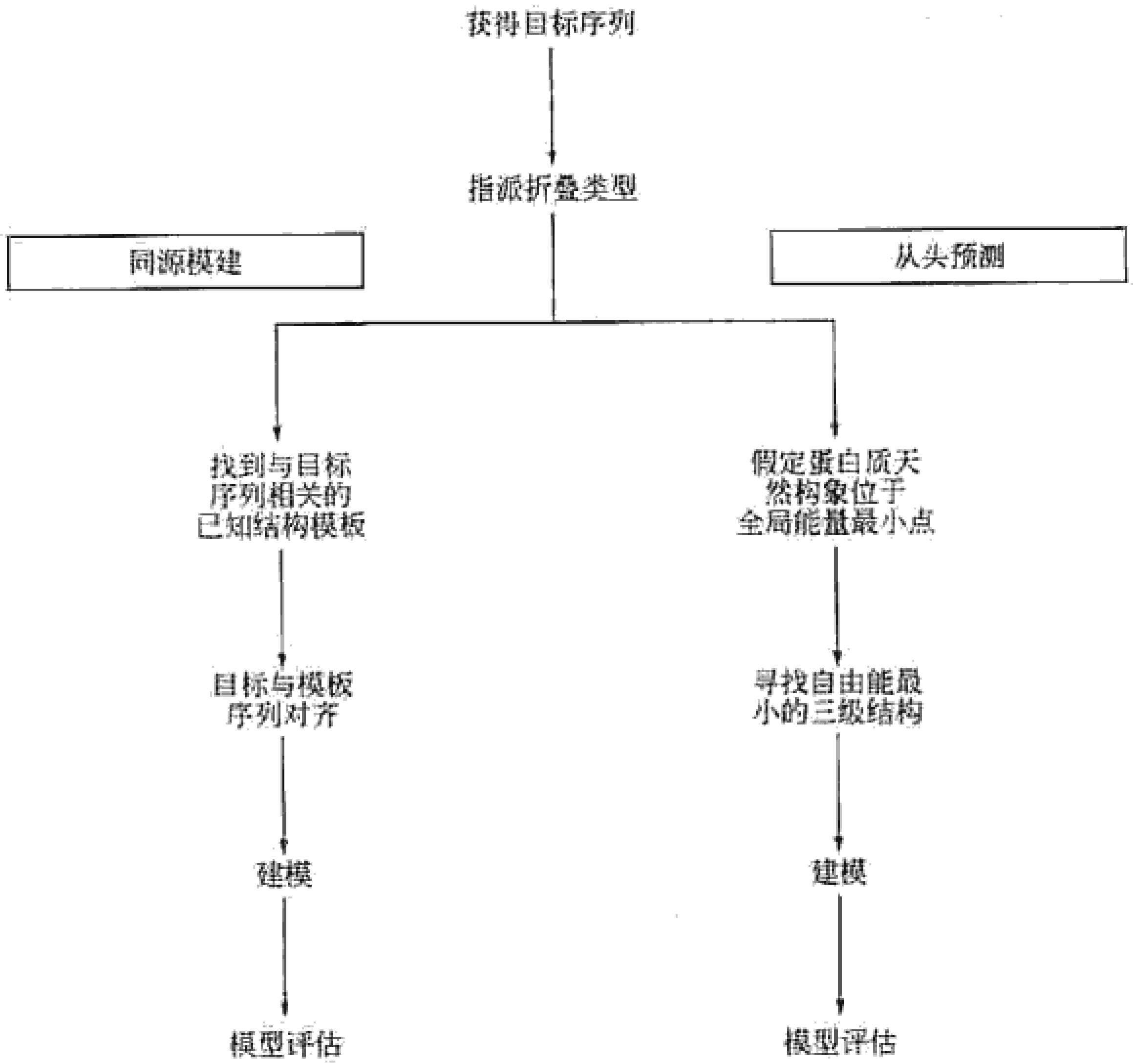


图 9-35 预测蛋白质结构的方法（Baker 和 Sali，2001）。当目标序列与某个已知结构蛋白具有同源迹象时，同源模建是最有效的方法。当同源性不足时，可以用从头预测的方法进行结构预测

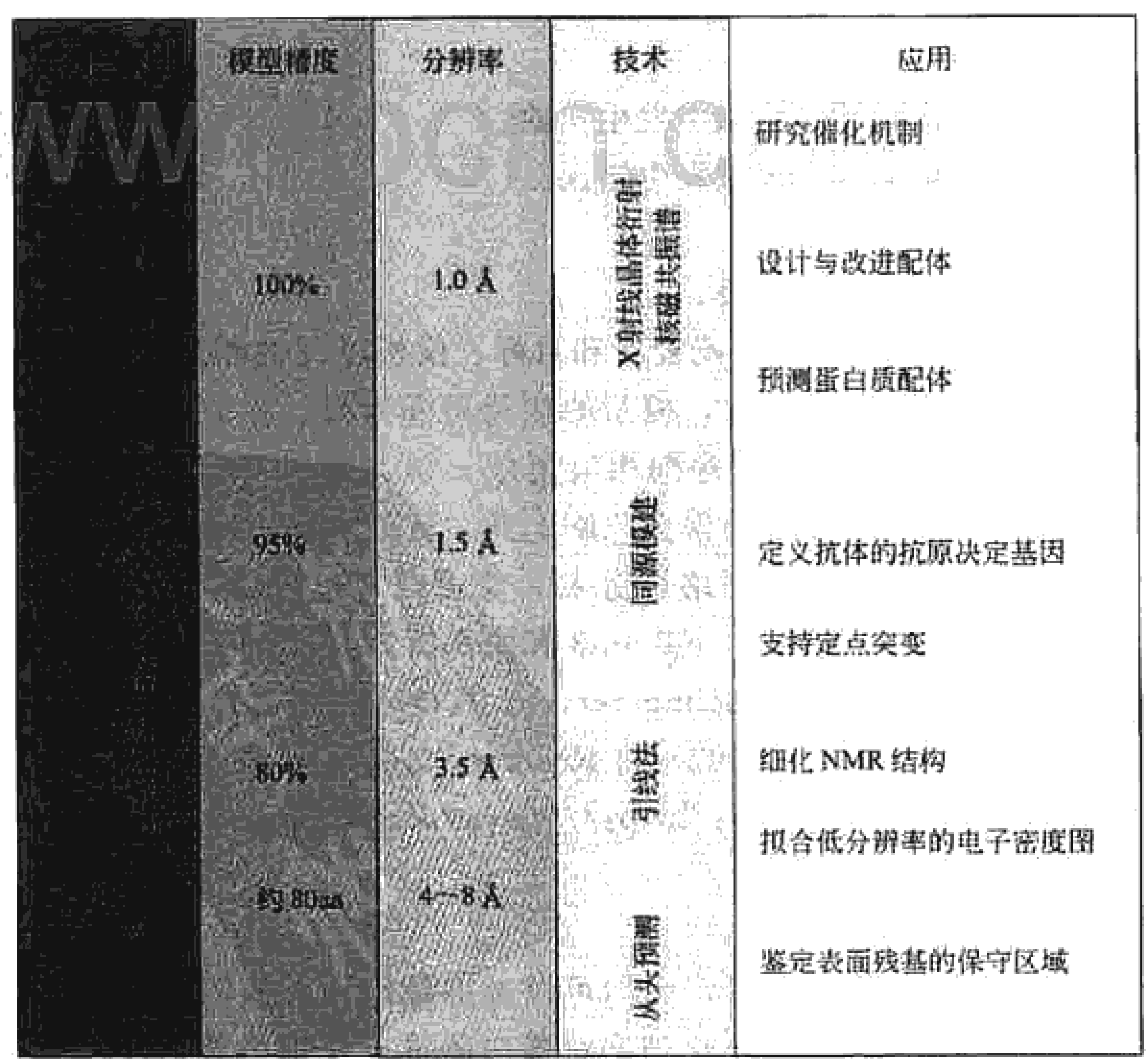


图 9-36 蛋白质结构预测精度和目标–模板序列相似度关系（来自 Baker 和 Sali，2001）

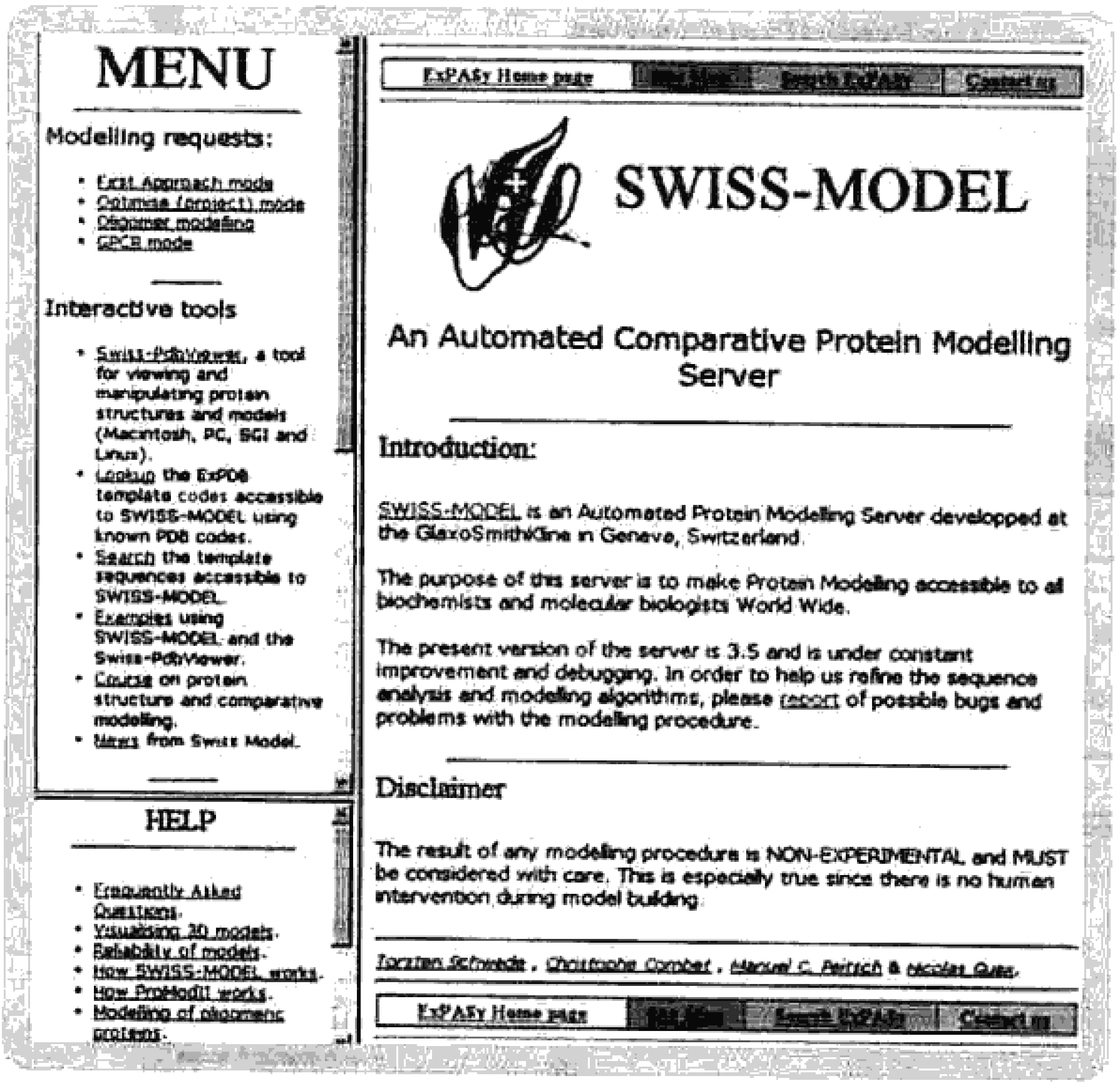


图 9-37 ExPASy SWISS-MODEL 是一个相当有用的同源模建服务器

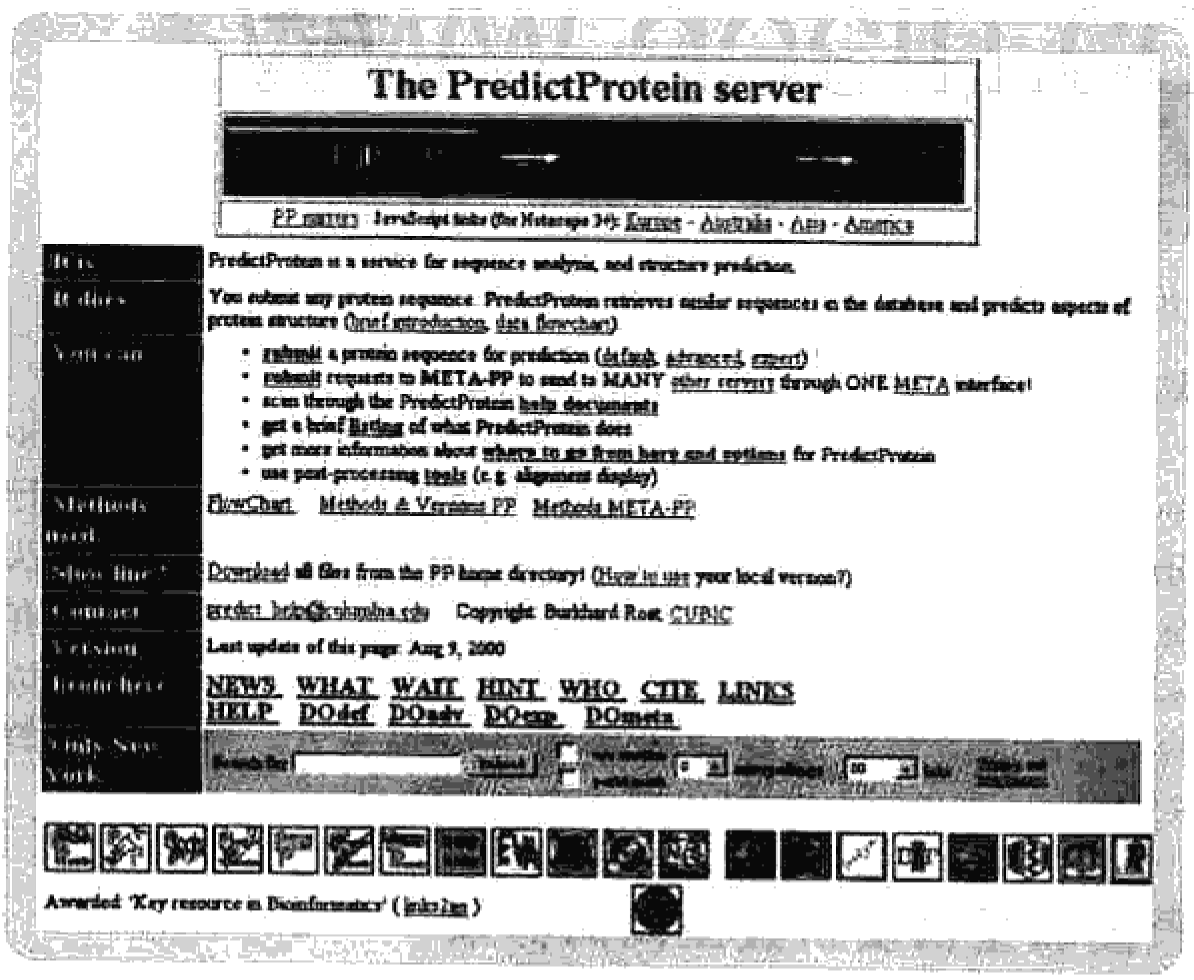


图 9-38 位于 http://cubic.bioc.columbia.edu.cn/prediction/的 Predict Protein 服务器提供同源模建的服务，在全球有大约 20 多个镜像。PredictProtein 提供数据库搜索、二级结构、残基的溶剂可及性、跨膜螺旋、蛋白质形状和其他特征的预测

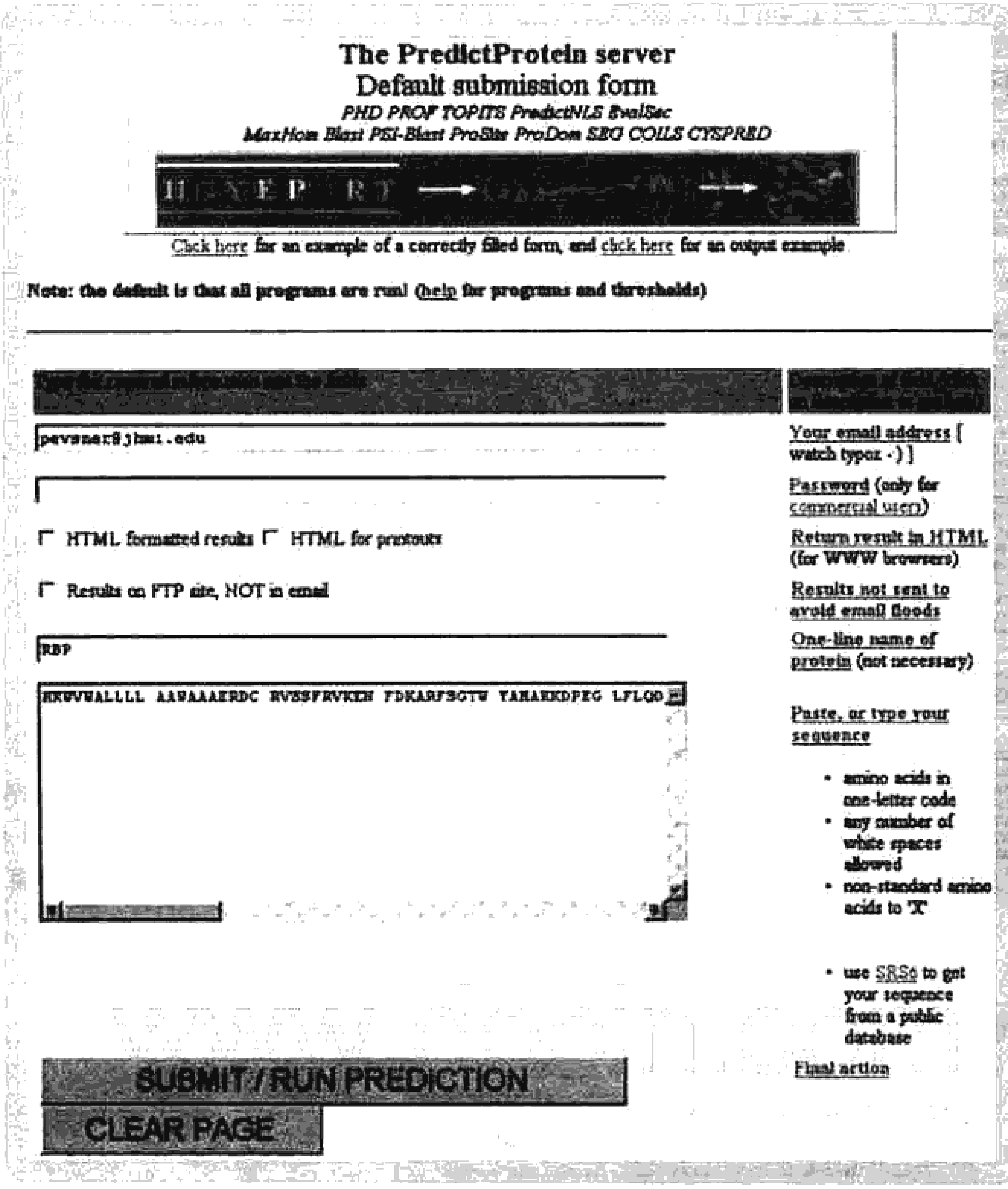


图 9-39 向 Predict Protein 服务器提交蛋白质查询的方式相当简单

> Wolf等（1999）联合了PSI-BLAST搜索得到的来自13个全基因组（包括细菌、真菌、真核生物）的蛋白质和来自SCOP的结构数据，对所有蛋白质中的20%～30%进行了折叠类型的预测。

型的有用方法。

Rosetta Stone模型是一个最为成功的从头预测的方法（Simons等，2001），它以9个氨基酸残基为单位对目标蛋白进行分析，将这些片段与PDB内已知结构比较，从而推断出整个肽链的结构。Bonneau等（2002）利用Rosetta Stone方法对所有未知结构的Pfam-A序列家族（第10章）进行建模。用已知结构进行检验，他们估计的准确率在60%左右，最好模型的RMSD<0.6nm。

9.4 蛋白质结构预测和蛋白质折叠空间的界限

我们能否鉴别出自然界的所有折叠类型？据估计，自然界折叠类型的数目在1000～

5000之间（Brenner等，1997；Burley，2000）。PDB收录了大约700种不同的蛋白质折叠类型，有一些为很多蛋白家族所共有，如TIM barrel（图9-23）在超过20种蛋白家族中出现。平均来说，每一个蛋白质有两个结构域，每一个结构域有（153±87）个残基（Orengo等，1999）。随着越来越多的蛋白质结构被解析，特殊折叠类型出现的比例在减少，因此可以期望，我们正逐步接近覆盖整个蛋白质序列空间的目标。

CASP（critical assessment of techniques for protein structure prediction），顾名思义，是世界范围内蛋白质结构预测技术与方法的竞赛。该竞赛每两年举行一次，一般第一年预测，第二年给出全面的评估。每届有数十支来自全球的队伍参与，每队的目的在于准确地预测一组目标序列的结构。目标序列的结构是已知的，但是没有公布。CASP组织者对预测结果进行评价（他们并不知道各个参赛队的信息）。在2000年10月的CASP4中，最多的错误来源于不正确的序列比对。

可从http://predictioncenter.llnl.gov获得关于CASP的信息。在CASP4中，目标蛋白包括了猪的β-乳球蛋白（β-lactoglobulin）（Target T123）。

9.5 蛋白质结构与疾病

氨基酸序列决定了蛋白质的三维结构，有时即使是一个残基的变化也会引起结构的显著改变。囊性纤维化病的病因是在编码囊性纤维化跨膜调节蛋白（CFTR）的基因内发生了变异（Ratjen和Doring，2003）。比较普遍的变异是ΔF508导致了CFTR508位苯丙氨酸的缺失。这种缺失的后果是改变了该蛋白质α螺旋的含量。CFTR正常情况下位于肺上皮细胞质膜，而这种结构变化在一定程度上阻碍了CFTR通过旁分泌到达此位点的过程。

你可以在LocusLink(LocusID:3043)获得关于血红蛋白链的一个简短定义，同时还可以找到指向OMIM的链接，在那里你可以得到由血红蛋白基因突变引发疾病的详细解释。我们将在第18章讨论镰刀形细胞贫血。

与疾病关联的蛋白质序列变化并不一定导致蛋白质结构上的巨大变化，这样的一个例子是镰刀型细胞贫血症（一种常见的遗传性血液疾病，它是由位于11号染色体短臂15.4位置上编码血红蛋白β链（也叫β球蛋白）的基因突变引起的。成人的血红蛋白在血液中可以将来自肺部的氧气分子携带到身体其他组织。正常的谷氨酸被缬氨酸替换后，在β球蛋白表面形成一个疏水的斑块，导致了多个血红蛋白分子的聚集。

表9-5列举了蛋白质序列细微变化导致结构变化而引起疾病的例子，包括了CFTR和β球蛋白。我们将在本章末尾研究它们的正常结构与变异结构。

表9-5 序列细微变化导致结构变化而引起疾病的例子

疾病	基因/蛋白	位点	参考序列
囊性纤维化病	CFTR	1080	NP_000483
镰刀形细胞贫血	血红蛋白β链	3043	NP_000509
疯牛病(BSE)	朊蛋白	5621	NP_000302
Alzheimer's症	淀粉样蛋白前体	351	NP_000475

注：CFTR表示囊性纤维化跨膜调节蛋白，BSE表示牛脑海绵状病。

9.6 展望

结构基因组学的目的在于定义蛋白质折叠空间中的所有结构。它与人类基因组计划有很

多相似之处，两者都需要全世界实验室的通力合作，都积累了大量的原始数据，而且数据库的规模都是呈指数形式增长。

如前所述，每年得到的新结构中特殊折叠类型出现的比例在减少，因此可以期望，在不久的将来我们能定义大部分蛋白质折叠类型。与此同时，下面一些问题需要我们给出解释：

① 蛋白质采用了有限数目的折叠类型。

② 一个折叠类型可以被不同的蛋白质采用，来执行完全不同的生物学功能。

③ 有着相同的生物学功能的蛋白质可以采用不同的折叠类型。

9.7 常见问题

生物学中一个最不可思议的问题是线性氨基酸序列是如何快速折叠成正确的三维结构的。对计算生物学家，仅从序列去预测三维结构相当困难，因此从头预测的方法远没有同源模建技术那么成功。

9.8 网络资源

网络资源见表 9-6。

表 9-6 链接到 PDB 的部分蛋白质结构数据库

数据库	评注	网站地址
3dee	结构域定义	http://jura.ebi.ac.uk:8080/3Dee/help/help_intro.html
BMCD	生物大分子晶体信息	http://wwwbmcd.nist.gov:8080/bmcd/bmcd.html
CATH	蛋白折叠分类	http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html
CE	完整与代表性的 PDB 结构比对	http://cl.sdsc.edu/ce.html
DSSP	二级结构分类	http://www.cmbi.kun.nl/swift/dssp/
Enzyme Structures Databases	酶的分类及术语	http://www.biochem.ucl.ac.uk/bsm/
FSSP	结构相似家族	http://www.ebi.ac.uk/dali/fssp
GRASS	图表描述与分析	http://honiglab.cpmc.columbia.edu/cgi-bin/GRASS/surfserv_enter.cgi
HSSP	同源二级结构	http://www.cmbi.kun.nl/swift/hssp
MMDB	三维结构数据库	http://www.ncbi.nlm.nih.gov/structure/
MEDLINE	从 NCBI 可直接访问 MEDLINE	http://www.nlm.nih.gov/databases/medline.html
NDB	核酸三维结构数据库	http://ndbserver.rutgers.edu/
PDBSum	蛋白质结构信息	http://www.biochem.ucl.ac.uk/bsm/pdbsum/
SCOP	结构分类	http://scop.mrc-lmb.cam.ac.uk/scop/
Tops	蛋白质结构模体的比较和结构拓扑图示	http://tops.ebi.ac.uk:80/tops/
VAST	NCBI 的结构比对工具	http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml
Whatcheck	蛋白质结构检验	http://biotech.ebi.ac.uk:8400/

注：改编自 Berman 等（2000）和 PDB（http://rcsb.org/pdb/links.html）。

9.9 问题讨论

1. PDB 是蛋白质结构数据的中心，其他诸如 SCOP 和 CATH 之类的数据库提供了哪些 PDB 所缺乏的信息？

2. 一个普遍的规律是：蛋白质结构比序列进化慢，因此两个蛋白质虽然序列上相似性有限，但还是有可能具有类似的结构［一个很好的例子是，一些载脂蛋白（如视黄醇结合蛋

白、增味剂结合蛋白和β-乳球蛋白）仅有很低的序列一致性，却具有极其相似的结构]。但是对于这个普遍规律还是有一些是例外的，你能找出一个吗？

9.10 问题

1. 观察一个蛋白质的结构。

① 选择一种载脂蛋白，去 NCBI 查找并观察其结构。你可以由 NCBI 主页进入 Structure 页面，然后进行查询；或者也可以直接利用 Entrez 搜索，选择“limits”，将结果限制在 PDB 内。如果你选择的是增味剂结合蛋白（odorant-binding protein，OBP），你得到的查询结果将包含了来自很多物种的 OBP。

② 在 MMDB 网页选择“View 3D Structure”，仔细研究这个页面上的所有链接。点击“View/Save Structure”（可能你先得下载 Cn3D 软件）。

③ 有两个窗口（即 Cn3D viewer 和 1D-viewer）会打开：注意它们是怎么联系的。改变 Cn3D viewer 的“style”选项，确认这个蛋白质的 α 螺旋和 β 片层区域。

2. 比较两种载脂蛋白的结构。

① 来到关于 1OBP 的 MMDB 页面，选择“Structure neighbor”（鼠标点击“protein”对应的图），打开的就是 NCBI 的 VAST 页面。列表中是与 OBP 相关的蛋白质，通过复选框选择其中的一个或两个（如 β-乳球蛋白和视黄醇结合蛋白），观察多重比对结果。

② 注意，会有两个窗口（Cn3D 和 DDV）打开，检查这两个工具之间的联系。你比较的几个蛋白质之间有什么共同点？区别又在哪里？在两个查看器中将保守区的氨基酸残基用高亮显示，指出保守 GXW 模体的位置。

3. 血红蛋白的 β 链的变异会引起镰刀形细胞贫血等一些疾病。设法从下列资源中找到野生型和突变型血红蛋白的 PDB 获取码：

① NCBI 的 Structure 页面；

② PDB 数据库；

③ CATH 或 SCOP 数据库；

④ 在 NCBI 网站上用 blastp 对 PDB 数据库进行搜索。

4. 镰刀形细胞贫血症是由 HBB、E6V 的点突变引起的（第六位的谷氨酸被替换为缬氨酸）。这种突变的后果是，血红蛋白四聚体聚集成块，使得整个红细胞发生形变，呈现为镰刀形。野生型血红蛋白的 PDB 标志码为 1HBB 或 4HBB，而突变型蛋白的标志码为 2HBS。利用 NCBI 的 VAST 工具对二者结构进行比较，第六位的谷氨酸位于蛋白质表面还是内部？谷氨酸→缬氨酸的突变引起了蛋白质的二级结构和三级结构的哪些变化。

5. ΔF508 是囊性纤维化病的一个主要病因，这个发生在编码囊性纤维化跨膜调节蛋白（CFTR）基因内的缺失导致了 508 位苯丙氨酸的缺失。Massiah 等（1999）利用核磁共振谱技术测定了长度为 26 个残基的野生型 CFTR 片段（PDB 获取码 1CKY）和长度为 25 个残基（缺少苯丙氨酸）的突变形片段（PDB 获取码 1CKZ）。这两种肽段因太小而无法用 NCBI 的 VAST 程序来进行结构比较。利用其他程序预测这两种肽段的二级结构的区别。

9.11 自测题

1. 比较两个远缘同源的蛋白：

(a) 相对于序列，它们倾向于具有更相似的三维构象；

(b) 相对于三维构象，它们倾向于具有更相似的序列；

(c) 它们的序列相似性与结构相似性都要达到一定程度；

(d) 并不能归纳出一个合理的结构和序列相似程度阈值

2. 典型的蛋白质二级结构预测算法通常计算一个蛋白质形成（ ）的概率

(a) α 螺旋；

(b) α 螺旋和 β 片层；

(c) α 螺旋、β 片层和卷曲；

(d) α 螺旋、β 片层，卷曲和多聚体

3. 相对于核磁共振谱，X 射线晶体衍射技术测定蛋白质结构的优点在于

(a) 能更方便地测定跨膜蛋白结构；

(b) 长晶体比制备核磁样品简单；

(c) 衍射数据更容易分析；

(d) 能够测定较大的蛋白质结构

4. PDB 数据库

(a) 是全世界最主要的大分子二级结构数据库；

(b) 其蛋白质结构数目与 SwissProt/TrEMBL 数据库中的序列数目相当；

(c) 包括了蛋白质，蛋白质–核酸和糖类的结构数据；

(d) 由 NCBI 和 EBI 共同管理

5. NCBI 的 VAST 程序

(a) 是一个利用 threading 方法观察蛋白质结构的浏览器工具；

(b) 是一个结构可视化工具，最多可同时查看两个蛋白质结构；

(c) 可以用来根据已知结构的蛋白质（具有 PDB 获取码）对 NCBI 的结构数据库进行搜索，但是不适用于未确认的结构；

(d) 可以对 NCBI 所有的结构数据库记录进行搜索，提供目标近邻结构的列表

6. Cn3D 是 NCBI 的一个分子结构查看器，其特点是

(a) 一个与自动的同源模建联系的菜单驱动程序；

(b) 一个命令行接口，对于结构分析十分有用；

(c) 一个结构查看器，同时伴有一个序列查看器；

(d) 一个可以观察立体结构图像的结构查看器

7. CATH 数据库对蛋白质结构进行了层次分类，前三个层次：类（class，C），构架（architecture，A）和拓扑（topology，T）描述的都是

(a) 蛋白质三级结构（三级结构的组成、堆垛、形状、方向和连接）；

(b) 蛋白质二级结构（二级结构的组成、堆垛、形状、方向和连接）；

(c) 蛋白质结构域的结构；

(d) 蛋白质超家族是根据同源结构域划分的

8. 同源模建与从头预测方法的区别在于

(a) 同源模建需要建立一个模型；

(b) 同源模建需要将目标与模板进行比对；

(c) 同源模建适用于所有蛋白质序列；

(d) 同源模建的准确度和目标、模板序列间的相似程度无关

9. 你想快速预测出一个蛋白质序列的结构。对目标序列进行 BLAST 和 PSI-BLAST 后，你发现在所有相关的已知结构的蛋白质中，最高的序列相似度有 17%，你将选择下面哪种方法？

(a) X 射线晶体衍射技术；

(b) 核磁共振谱技术；

(c) 将序列提交到蛋白质结构预测服务器进行同源模建；

(d) 将序列提交到蛋白质结构预测服务器进行从头预测

9.12 推荐读物

有很多关于结构基因组学和蛋白质结构预测的优秀综述，如 Janet Thornton，Christine Orengo 及其合作者们（Thornton 等，2000；Todd 等，2001；Orengo 等，2001）；Stephen Burley（Burley 等，1999；Burley，2000）；Brunel 大学的 David T. Jones（Jones，2001）；还有其他（Teichman 等，1999；Holm，1998；Holm 和 Sander，1996b；Domingues 等，2000）文章。大多数文章讨论了蛋白质结构和功能的关系，对蛋白质结构预测的各种方法进行了评估。另外推荐在"Nature Structure Biology"增刊上关于结构基因组学的一篇综述（Mark Gerstein，2000）(http://strutbio.nature.com)。关于同源模建，请参考 Andrej Sali 与其合作者（Marti-Renom 等，2000）的综述。

还有不少文章对本章中提到的数据库有详细的介绍。关于 PDB，请参考 Berman 等的文章（Berman 等，2000；Westbrook 等，2002）。关于 CATH，请参照 Christine Orengo 等的综述（Pearl 等，2000）。Murzin 等的文章中对 SCOP 有介绍；关于 Dali 和 FSSP，请参考 Dietmann 等（2001）以及 Holm 与 Sander 的文章（1996a）。

参考文献

Baker, D., and Sali, A. Protein structure prediction and structural genomics. *Science* **294**, 93-96 (2001).

Berman, H. M., et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242 (2000).

Berman, H. M., et al. The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* **58**, 899-907 (2002).

Bonneau R., Strauss C. E., Rohl C. A., Chivian D., Bradley P., Malmstrom L., Robertson T., Baker D. De novo prediction of three-dimensional structures for major protein families. *J. Mol. Biol.* **322**, 65-78 (2002).

Brenner, S. E., Chothia, C., and Hubbard, T. J. Population statistics of protein structures: Lessons from structural classifications. *Curr. Opin. Struct. Biol.* **7**, 369-376 (1997).

Brenner, S. E. Target selection for structural genomics. *Nat. Struct. Biol.* **7** (Suppl.), 967-969 (2000).

Burley, S. K., et al. Structural genomics: Beyond the human genome project. *Nat. Genet.* **23**, 151-157 (1999).

Burley, S. K. An overview of structural genomics. *Nat. Struct. Biol.* **7** (Suppl.), 932-934 (2000).

Bystroff, C., and Shao, Y. Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. *Bioinformatics.* **18** Suppl 1, S54-61 (2002).

Chang, G., and Roth, C. B. Structure of MsbA from E. coli: A Homolog of the Multidrug Resistance ATP Binding Cassette (ABC) Transporters. *Science* **293**, 1793-800 (2001).

Chou, P. Y., and Fasman, G. D. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* **47**, 45-148 (1978).

Dietmann, S., et al. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res.* **29**, 55-57 (2001).

Domingues, F. S., Koppensteiner, W. A., and Sippl, M. J. The role of protein structure in genomics. *FEBS Lett.* **476**, 98-102 (2000).

Flower, D. R., North, A. C., and Sansom, C. E. The lipocalin protein family: Structural and sequence overview. *Biochim. Biophys. Acta* **1482**, 9-24 (2000).

Garnier, J., Gibrat, J. F., and Robson, B. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* **266**, 540-553 (1996).

Garnier, J., Osguthorpe, D. J., and Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97-120 (1978).

Gerstein, M. Integrative database analysis in structural genomics. *Nat. Struct. Biol.* **7** (Suppl.), 960-963 (2000).

Gerstein, M., and Levitt, M. A structural census of the current population of protein sequences. *Proc. Natl. Acad. Sci. USA* **94**, 11911-11916 (1997).

Holm, L., and Sander, C. The FSSP database: Fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res.* **24**, 206-209 (1996a).

Holm, L., and Sander, C. Mapping the protein universe. *Science* **273**, 595-603 (1996b).

Holm, L. Unification of protein families. *Curr. Opin. Struct. Biol.* **8**, 372-379 (1998).

Jones, D. T. Protein structure prediction in genomics. *Brief. Bioinform.* **2**, 111-125 (2001).

Koonin, E. V., Wolf, Y. I., and Karev, G. P. The structure of the protein universe and genome evolution. *Nature* **420**, 218-223 (2002).

Marti-Renom, M. A., et al. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291-325 (2000).

Massiah, M. A., Ko, Y. H., Pedersen, P. L., and Mildvan, A. S. Cystic fibrosis transmembrane conductance regulator: Solution

structures of peptides based on the Phe508 region, the most common site of disease-causing DeltaF508 mutation. *Biochemistry* **38,** 7453–7461 (1999).

McGuffin, L. J., and Jones, D. T. Targeting novel folds for structural genomics. *Proteins* **48,** 44–52 (2002).

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540 (1995).

Orengo, C. A., et al. CATH—a hierarchic classification of protein domain structures. *Structure* **5,** 1093–1108 (1997).

Orengo, C. A., Bray, J. E., Hubbard, T., LoConte, L., and Sillitoe, I. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins* (Suppl.) **3,** 149–170 (1999).

Orengo, C. A., Sillitoe, I., Reeves, G., and Pearl, F. M. Review: What can structural classifications reveal about protein evolution? *J. Struct. Biol.* **134,** 145–165 (2001).

Osguthorpe, D. J. Ab initio protein folding. *Curr. Opin. Struct. Biol.* **10,** 146–152 (2000).

Pearl, F. M., et al. Assigning genomic sequences to CATH. *Nucleic Acids Res.* **28,** 277–282 (2000).

Pearson, W. R., and Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448 (1988).

Przybylski, D., and Rost, B. Alignments grow, secondary structure prediction improves. *Proteins* **46,** 197–205 (2002).

Ratjen, F., and Döring, G. Cystic fibrosis. *Lancet* **361,** 681–689 (2003).

Rost, B., and Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232,** 584–99 (1993a).

Rost, B., and Sander, C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sci. USA* **90,** 7558–7562 (1993b).

Rost, B. Review: Protein secondary structure prediction continues to rise. *J. Struct. Biol.* **134,** 204–218 (2001).

Shortle, D. Prediction of protein structure. *Curr. Biol.* **10,** R49–51 (2000).

Simons, K. T., Strauss, C., and Baker, D. Prospects for ab initio protein structural genomics. *J. Mol. Biol.* **306,** 1191–1199 (2001).

Taylor, W. R., and Orengo, C. A. Protein structure alignment. *J. Mol. Biol.* **208,** 1–22 (1989a).

Taylor, W. R., and Orengo, C. A. A holistic approach to protein structure alignment. *Protein Eng.* **2,** 505–519 (1989b).

Teichmann, S. A., Chothia, C., and Gerstein, M. Advances in structural genomics. *Curr. Opin. Struct. Biol.* **9,** 390–399 (1999).

Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N., and Orengo, C. A. From structure to function: Approaches and limitations. *Nat. Struct. Biol.* **7** (Suppl.), 991–994 (2000).

Todd, A. E., Orengo, C. A., and Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307,** 1113–1143 (2001).

Venclovas, C., Zemla, A., Fidelis, K., and Moult, J. Comparison of performance in successive CASP experiments. *Proteins* (Suppl.), **5,** 163–170 (2001).

Wang, Y., et al. MMDB: 3D structure data in Entrez. *Nucleic Acids Res.* **28,** 243–245 (2000).

Westbrook, J., et al. The Protein Data Bank: Unifying the archive. *Nucleic Acids Res.* **30,** 245–248 (2002).

Wolf, Y. I., Brenner, S. E., Bash, P. A., and Koonin, E. V. Distribution of protein folds in the three superkingdoms of life. *Genome Res.* **9,** 17–26 (1999).

Wood, T. C., and Pearson, W. R. Evolution of protein sequences and structures. *J. Mol. Biol.* **291,** 977–95 (1999).

Zemla, A., Venclovas, Moult, J., and Fidelis, K. Processing and evaluation of predictions in CASP4. *Proteins* (Suppl.), 13–21 (2001).

Tabelle 2

VERGLEICH EINER HOMOLOGEN REGION IN CYTOCHROM-C VERSCHIEDENER HERKUNFT

Rind	...-Val-Glu(NH$_2$)-Lys-CyS-Ala-Glu(NH$_2$)-CyS-His-Thr-Val-Glu-Lys-...
Pferd	................-Lys-CyS-Ala-Glu(NH$_2$)-CyS-His-Thr-Val-Glu-Lys-...
Schwein	................-Lys-CyS-Ala-Glu(NH$_2$)-CyS-His-Thr-Val-Glu-Lys-...
Lachs	...-Val-Glu(NH$_2$)-Lys-CyS-Ala-Glu(NH$_2$)-CyS-His-Thr-Val-Glu- ......
Huhn	...-Val-Glu(NH$_2$)-Lys-CyS-*Ser*-Glu(NH$_2$)-CyS-His-Thr-Val-Glu-.......
Seiden-spinner	...-Val-Glu(NH$_2$)-*Arg*-CyS-Ala-Glu(NH$_2$)-Cys-His-Thr-Val-Glu-.......
Hefe	...-*Phe-Lys-Thr*-------- *Arg*--CyS-*Glu*--*Leu*--------CyS-His-Thr-Val-Glu-.......

als mit der in Säugetier-Cytochromen ermittelten identisch, in Hühner-Cytochrom-c hingegen war an die Stelle eines Alanin-Restes ein Serin-Rest getreten. Ein Invertebraten-Cytochrom-c, das des Seidenspinners Bombyx

(a)

www.docin.com

Tabelle 3

ARTUNTERSCHIEDE IN INSULIN (NACH HARRIS, SANGER UND NAUGHTON, 1956)

Rind	...-CyS-*Ala-Ser-Val*-CyS-...
Schwein	...-CyS-*Thr-Ser-Ileu*-CyS-...
Schaf	...-CyS-*Ala-Gly-Val*-CyS-...
Pferd	...-CyS-*Thr-Gly-Ileu*-CyS-...
Wal	...-CyS-*Thr-Ser-Ileu*-CyS-...

(b)

由于很多蛋白序列早在20世纪五六十年代就被测定，人们很自然地会想到对它们进行比较。(a) Hans Tuppy (1958) 将来自牛 (beef)、马 (horse)、猪 (pig)、大麻哈鱼 (salmon)、鸡 (chicken)、蚕 (silk-worm)、酵母 (yeast) 的细胞色素C进行了比对。结果显示引入空白 (gap) 之后，像酵母与哺乳动物这样远缘的序列还是可以很好的对齐；(b) Tuppy (1958) 还对来自牛 (beef)、马 (horse)、猪 (pig)、羊 (sheep)、鲸鱼 (whale) 的胰岛素序列进行了比对。在这种情况下，他发现两个半胱氨酸之间的区域缺乏序列保守性

第 10 章

多序列比对

10.1 引言

当我们在研究一个蛋白质或基因时，经常会考虑这样一个很基本的问题：它与其他蛋白质的关系如何。生物学中的序列通常以家族的形式存在，这些家族由同一个体内的相关基因（paralogs），同一种群（polymorphic variants）或不同种群的基因（orthologs）组成。基因组内发生重复（duplications）或由物种形成而出现直系同源（orthologs）等原因导致了序列的分歧。我们已经学习了蛋白质、核酸的两两比对（第 3 章），同时我们看到了可以由 BLAST 或其他数据库搜索得到的多个相关序列（第 4 章、第 5 章）。在蛋白质结构域的研究过程中（第 8 章），我们可以发现结构数据库通常以多序列比对的形式来显示同源蛋白序列（第 9 章）。在本章中，我们将考虑多序列比对中存在的一般问题。尽管可以对蛋白质、DNA 序列进行多重比对，但是很多数据库的比对只针对蛋白质家族。

由于同源序列通常保持了相似的结构和功能，因而多序列比对就显得很有意义。通过序列的多重比对，我们可以确定一个基因或蛋白家族的成员特征。如果我们知道其中一个蛋白的一些特征（如 RBP4 转运疏水配体），在确认了一些与之同源的蛋白后，我们可以预测它们可能行使相似的功能。绝大多数蛋白质是通过基因组 DNA 与 cDNA 测序鉴定的，因此大多数蛋白质的功能是基于同源分析确定，而非实验方法。

10.1.1 多序列比对的定义

蛋白家族的特征是用存在一组同源序列的多重比对来定义的。一个多重比对就是一组可以部分或整体对齐的蛋白质或核酸序列（3 个或 3 个以上）。相同或相似的氨基酸残基排在同一列上，这些对齐的残基在进化意义上是同源的：来自共同的祖先。并且我们还假定从结构的角度来看，这些残基也是同源的：在三维结构中，对齐的残基也倾向于占据对应的位置。

对于关系很近的一组序列，很容易产生多序列比对，甚至可以直接观察得到（我们见过这样的例子，图 3-7 和图 3-8，GAPDH 和 RBPs）。当序列间出现一些分歧时，多序列比对过程中出现的问题就很难解决了，如空白数量和位置的估计就相当困难。图 3-9 所示的一些

载脂蛋白比对就是这样的一个例子。在本章中我们将比较 5 个远缘载脂蛋白和 5 个近缘 RBP 直系同源物的多重比对结果。

通常情况下，序列比结构具有更快的进化速度。比较 RBP4 和 β-乳球蛋白（β-lactoglobulin），我们可以发现它们之间的序列相似性只有 26%（图 3-5），而结构却几乎一样（图 3-1 和图 9-18）。在产生多序列比对的时候，我们几乎不可能找到所有应该对齐（结构角度）的氨基酸。通常我们没有高分辨率的结构，也缺少用以分辨结构域的功能信息（例如形成酶活性位点的特殊氨基酸），因而只能依赖序列信息进行多重比对。在结构已知的情况下，先进行多重序列比对，然后用结构信息加以验证的方法也是可行的。对于有一定分歧但是明显相关的一对蛋白序列（序列相似性大于 30%），Chothia 和 Lesk（1986）发现大约有 50% 的氨基酸残基在两个结构中的位置是重叠的。

可以根据下面的特征来确定是否对齐某些氨基酸残基。

① 一些高度保守的残基（如参与形成二硫键的半胱氨酸）。

② 形成保守模体（motif），如跨膜结构域和免疫球蛋白结构域。在第 8 章中我们见过蛋白质结构域和模体的例子。

③ 蛋白质二级结构的保守特征，如参与形成 α 螺旋、β 片层和可变区的残基。

④ 显示出一致插入或缺失模式的区域。

10.1.2 多序列比对的典型应用和实际策略

什么时候使用和为什么使用多重比对？

① 如果所研究的蛋白质或基因与另一组蛋白质有联系，那么这些蛋白质可以提供可能的功能、结构、进化方面的信息。

② 大多数蛋白质家族中有远缘的成员。与两两比对相比，多序列比对能够更敏感地发现同源关系。

③ 在检查某次数据库搜索结果（如 BLAST 搜索）时，多重比对形式的结果能更容易显示保守残基与模体。

④ 如果研究 cDNA 克隆，按照惯例你可能会对它们进行测序。多序列比对可以显示结果中是否有矛盾之处。

⑤ 分析物种数据可以揭示很多生物学问题（如进化、结构和功能等方面）。Entrez 的 PopSet 部分（下文会提到）包含了核酸（和蛋白质）的物种数据集，可以多重比对的形式显示。

⑥ 当一个物种的基因组被完整测序，数据分析的一个主要部分是定义所有基因产物所归属的蛋白家族。数据库搜索进行高效的多重比对，将每一个新蛋白（或基因）与其他所有家族的蛋白质进行比较。

⑦ 在第 11 章中我们将看到如何利用多序列比对数据构建系统发生树。建树的一个最关键的步骤是产生最佳比对。

> 可以从下面地址访问 ClustalW：http://www2.ebi.ac.uk/clustalw/。

⑧ 很多基因的调节区含有转录因子结合的共有序列。

10.1.3 Feng 和 Doolittle 的渐进比对方法

在目前发展的众多多序列比对方法中，最常用的是来自 Da-Fei Feng 和 Russell Doolittle（1987，1990）的渐进比对。之所以称为渐进比对，是因为这种方法需要先对

所有的序列计算两两比对的分值，然后从关系最近的一对序列开始，逐步加入其他序列。

我们对从 Entrez 选出的 5 个载脂蛋白进行比对以演示这一过程。将这 5 个蛋白质序列输入 ClustalW 程序（图 10-1）——最流行的多序列比对程序之一（Thompson 等，1994）。

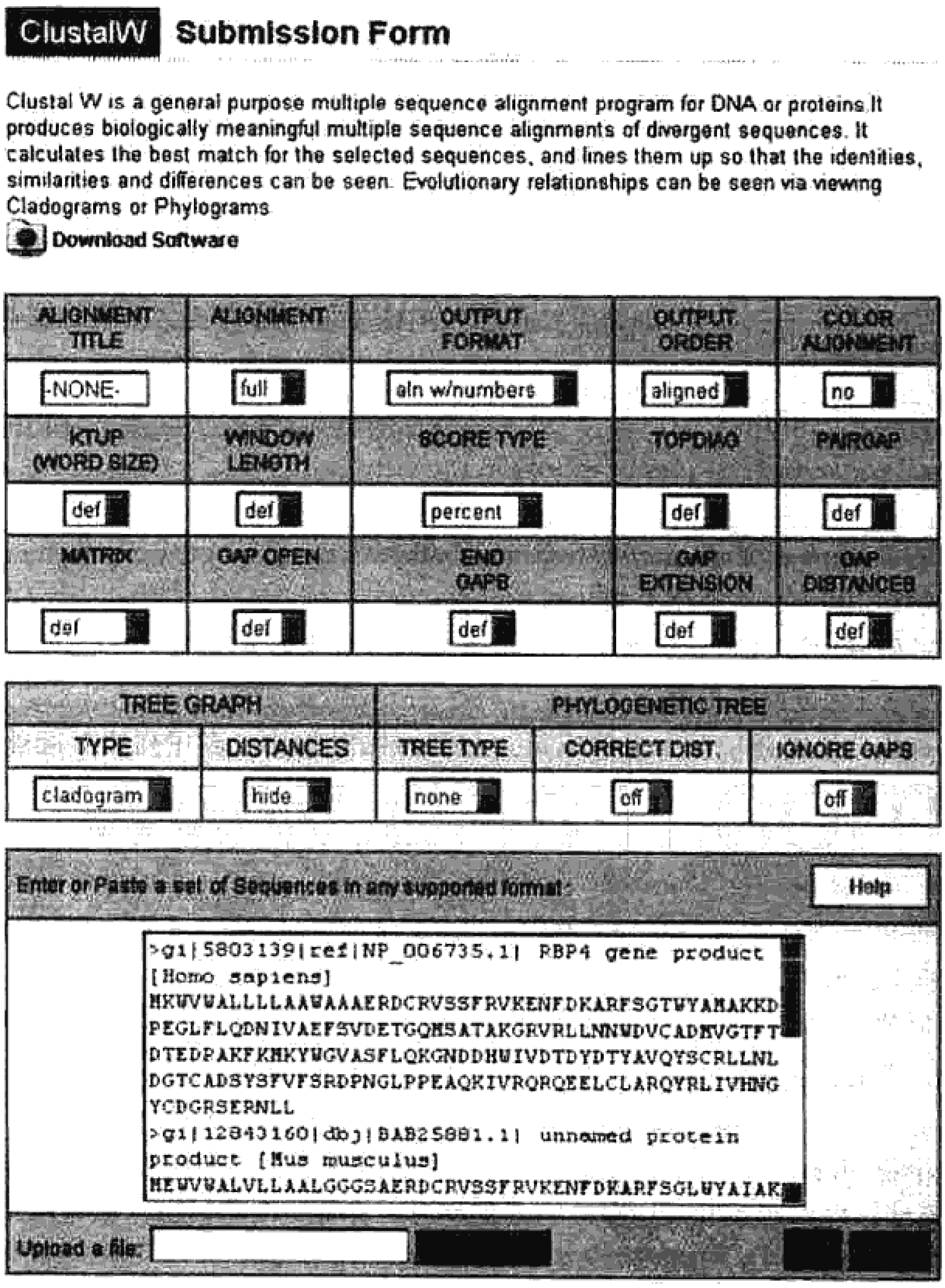


图 10-1 利用 EBI 的 ClustalW 服务器（http://www2.ebi.ac.uk/clustalw/）对 5 个载脂蛋白进行多重比对。这 5 个蛋白质是人类的视黄醇结合蛋白（retinol binding protein, NP_006735），老鼠的视黄醇结合蛋白（retinol binding protein, BAB25881），人类的脂蛋白 D（apolipoprotein D, NP_001638），老鼠的主要尿蛋白 4（major urinary protein 4, P11590），大肠杆菌的外膜脂蛋白 BLC（P39281）。每一个蛋白质序列均采用 Entrez 的 FASTA 格式

第一步，利用 Needleman 和 Wunsch（1970；第 3 章）的比对方法对所有的序列进行两两比对（图 10-2，阶段 1）。如图所示，对 5 个序列，要进行 10 次两两比对。

进行两两比对的算法可以产生原始的相似性分值。很多渐进序列比对算法（包括 ClustalW）使用距离矩阵（不是相似度矩阵）来描述蛋白质之间的关系。相似度分值到距离分值的换算见框 10-1。计算距离的目的在于产生辅助导向树，进而构建比对。

阶段1：产生一系列的两两比对

```
Sequence format is Pearson
Sequence 1: gi|5803139|ref|NP_006735.1|          199 aa
Sequence 2: gi|12843160|dbj|BAB25881.1|          201 aa
Sequence 3: gi|4502163|ref|NP_001638.1|          189 aa
Sequence 4: gi|127528|sp|P11590|MUP4_MOUSE       178 aa
Sequence 5: gi|732003|sp|P39281|BLC_ECOLI        177 aa

Start of Pairwise alignments
Aligning...
Sequences (1:2) Aligned. Score:  84 <---- 1
Sequences (1:3) Aligned. Score:  14
Sequences (1:4) Aligned. Score:  8
Sequences (1:5) Aligned. Score:  12
Sequences (2:3) Aligned. Score:  17
Sequences (2:4) Aligned. Score:  9
Sequences (2:5) Aligned. Score:  19
Sequences (3:4) Aligned. Score:  9
Sequences (3:5) Aligned. Score:  27
Sequences (4:5) Aligned. Score:  7
```

阶段2：计算距离矩阵，建立辅助导向树

```
(
(
(
gi|5803139|ref|NP_006735.1|:0.09385,
gi|12843160|dbj|BAB25881.1|:0.05691)
:0.34199,
gi|127528|sp|P11590|MUP4_MOUSE:0.48994)
:0.06118,
gi|4502163|ref|NP_001638.1|:0.35805,
gi|732003|sp|P39281|BLC_ECOLI:0.36511);
```



图 10-2 像 ClustalW 之类的比对程序都采用了 Feng 和 Doolittle 的渐进比对方法（1987）。在第一阶段，对五个远缘载脂蛋白进行两两比对（图 10-1）。注意最高的分值属于人类与老鼠的 RBP 比对（分值 84，箭头 1）。在第二阶段，根据五个序列的远近关系建立辅助导向树。利用 ClustalW 网络服务器上的 JalViewer 工具可以观察图形化的导向树。树的分支长度反映了序列之间的距离，注意和图 10-4 比较

框 10-1 相似性和距离度量

以树的形式来组织蛋白质或核酸序列通常可以表示序列之间的差异程度。度量距离的一种方法是计算两两比对中错配的数目。另一种方法（采用 Feng 和 Doolittle 的比对算法）是将相似性分值转化为距离分值。对所有要进行多重比对的蛋白序列进行一系列两两比对，得到相似性分值，然后利用下面的等式转化成距离分值：

$$D=-\ln S_{eff} \tag{10-1}$$

其中

$$S_{eff}=\frac{S_{real(ij)}-S_{rand(ij)}}{S_{iden(ij)}-S_{rand(ij)}}\times 100 \tag{10-2}$$

$S_{real(ij)}$ 指观测到的两个比对序列的相似性分值；$S_{iden(ij)}$ 指两个序列分别与自身比较后得分的平均值（如果 i 和 i 比较后的得分是 20，j 和 j 比较后的得分是 10，那么 $S_{iden(ij)}=15$）；$S_{rand(ij)}$ 是将序列打乱后，随机抽取得到（如 1000 次）的平均比对得分。S_{eff} 是一个归一化的分值，如果序列 i 和 j 没有相似性，则 $S_{eff}=0$，距离就是无穷。如果序列 i 和 j 一致，则 $S_{eff}=1$，距离是 0。

如果要对 N 个序列进行多重比对，最初两两比对的数目是 $(N-1)N/2$。$N=5$，则需要 10 次两两比对。

注意到在本例中，两两比对的最高分值属于人与鼠的 RBP4（见图 10-2，箭头 1）。我们下面会提到大部分多重比对算法要利用全局的两两比对，少数使用局部比对。对于关系很近的 RBP，两两比对得到的相似度分值都相当高（图 10-4）。

第二步，利用相似度矩阵（或距离矩阵）产生辅助导向树。有两种建立导向树的方法：算术平均不加权组队法（unweighted pair group method of arithmetic averages，UPGMA）和连接近邻方法（neighbor-joining method），我们将在第 11 章中给出这些算法的定义。树结构的两大特征是它的分支顺序和分支长度（与进化距离成比例）。因此导向树反映了所有蛋白序列间的进化关系。在 ClustalW 中，导向树是用一种语法来表示的（图 10-2 和图 10-4，第二步）。可以发现最小的距离值出现在人类与老鼠的 RBP 之间。利用 JalView 选项可以观察图形化的导向树（图 10-2 和图 10-4，第二步）。导向树并不等同于系统发生树，它只是为下一步在比对中添加序列提供参考。

第三步，根据导向树中的位置关系建立多重比对。具体算法是：先选择最接近的一个两两比对，接着在此基础上加入下一个最近序列形成三重比对或加入下一个最近的两两比对。重复此过程直到用完所有序列（图 10-3 和图 10-5，第三步）。

Feng-Doolittle 方法有这样一个规则："一旦是空白，永远是空白"（once a gap，always a gap）。首先是最近的序列被对齐，当较远的序列加入时，可能会在对齐时产生一些空白。这个规则的合理解释是在决定是否插入空白时，对距离较近的比对应优先考虑。

10.1.4 从多序列比对到隐马模型

渐进序列比对算法是与动态规划、两两比对紧密联系的，而隐马模型（hidden Markov models，HMM）是序列比对的另一个有力工具（Krogh，1998；Eddy，1998；Baldi 等，1994；Birney，2001）。它实际上是一个描述在每个位点氨基酸残基出现概率的概率模型。HMM 可以将一个多重比对转化为位点特异性打分系统。HMM 的一个普遍应用是在一个基于 HMM 的数据库（如 Pfam）中查询单个蛋白质序列。

考虑图 10-6(a) 中 5 个载脂蛋白 GXW 保守区的 5 个氨基酸残基，通过估计各种氨基酸

在这些位点出现的概率可以建立 HMM [图 10-6(b)]。从某种意义上说 HMM 与 PSI-BLAST（第 5 章）的位点特异性打分矩阵（PSSM）的计算过程类似。通过 HMM 的概率，对任意确定的一个查询序列可以计算出一个分值，如图 10-6(c) 所示的 GEWYE。HMM 是一种用状态来描述一个序列中各个位置的模型 [图 10-6(d)]。

相对于 PSSM，HMM 的图示更复杂。首先，通过多序列比对定义一系列概率，一个 HMM 模型的结构如图 10-7 所示（Krogh 等，1994；Krogh，1998）。最底层是一列主要状态（从“begin”到 m1－m5，再到“end”），这些状态对应一个序列的氨基酸残基（如 GTWYA）。第二行由插入状态组成（图 10-7，菱形图标 i1－i5）。这些状态模拟比对中的可变区域，允许在适当的地方进行插入。顶部第三行的圆形图标称为删除状态，它们对应了比对中的空白。

> HMM 包含了观察序列（由 HMM 模拟的残基序列）和隐含的状态序列（从观察序列中推测得到），这就是隐马模型名字的由来。
> HMMER 可从下面地址访问：Http://hmmer.wustl.edu，作者是 Sean Eddy，运行于 Unix 平台。

一个 HMM 的蛋白质序列是由一系列相互联系的状态来定义的（图 10-7，箭头）。每一个状态对匹配的氨基酸残基有一个概率分布。HMM 的观察序列就是多序列比对中一致的部分，同样还有描述马尔可夫链转移路径的状态序列。

Stage 3: progressively align the sequences following the branching order of the tree

```
CLUSTAL W (1.82) multiple sequence alignment

gi|5803139|ref|NP_006735.1|       MKWVWALLLLAAWA--AAERD------CRVSSFR----VKENFDKARFSG 38
gi|12843160|dbj|BAB25881.1|       MEWVWALVLLAALGGGSAERD------CRVSSFR----VKENFDKARFSG 40
gi|4502163|ref|NP_001638.1|       --MVMLLLLLSALAGLFGAAEGQAFHLGKCPNPP----VQENFDVNKYLG 44
gi|732003|sp|P39281|BLC_ECOLI     ---MRLLPLVAAATAAFLVVA------CSSPTPPRGVTVVNNFDAKRYLG 41
gi|127528|sp|P11590|MUP4_MOUSE    ----MKLLLCLGLTLVCIHAE-------EATSKG------QNLNVEKING 33
                                      * *  .                  ..        :*::  :  *

gi|5803139|ref|NP_006735.1|       TWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCAD 88
gi|12843160|dbj|BAB25881.1|       LWYAIAKKDPEGLFLQDNIIAEFSVDEKGHMSATAKGRVRLLSNWEVCAD 90
gi|4502163|ref|NP_001638.1|       RWYEIEK-IPTTFENGRCIQANYSLMENGKIKVLNQELR--ADGTVNQIE 91
gi|732003|sp|P39281|BLC_ECOLI     TWYEIARFDHRFERGLEKVTATYSLRDDGGLNVINKGYNP-DRGMWQQSE 90
gi|127528|sp|P11590|MUP4_MOUSE    EWFSILLASDKREK-IEEHGSMRVFVEHIHVLENSLAFKFHTVIDGECSE 82
                                   *: :                    :    . :   :                :

gi|5803139|ref|NP_006735.1|       MVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRL 138
gi|12843160|dbj|BAB25881.1|       MVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYDTFALQYSCRL 140
gi|4502163|ref|NP_001638.1|       GEATPVNLTEPAKLEVKFS---WFMP--SAPYWILATDYENYALVYSC-- 134
gi|732003|sp|P39281|BLC_ECOLI     GKAYFTGAPTRAALKVSFFG-------------PFYGGYNVIALDREYR- 126
gi|127528|sp|P11590|MUP4_MOUSE    IFLVADKTEKAGEYSVMYDG--------FNTFTILKTDYDNYIMFHLIN- 123
                                            .  .: :                   .  .*:   :

gi|5803139|ref|NP_006735.1|       LNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNG 188
gi|12843160|dbj|BAB25881.1|       QNLDGTCADSYSFVFSRDPNGLSPETRRLVRQRQEELCLERQYRWIEHNG 190
gi|4502163|ref|NP_001638.1|       TCIIQLFHVDFAWILARNPN-LPPETVDSLKNILTSNNIDVKKMTVTDQV 183
gi|732003|sp|P39281|BLC_ECOLI     HALVCGPDRDYLWILSRTPT-ISDEVKQEMLAVATREGFDVSKFIWVQQP 175
gi|127528|sp|P11590|MUP4_MOUSE    --EKDGKTFQLMELYGRKADLNSDIKEKFVKLCEEHGIIKENIIDLTKTN 171
                                         .   : .* .   .         :        :  .    .

gi|5803139|ref|NP_006735.1|       YCDGRSERNLL 199
gi|12843160|dbj|BAB25881.1|       YCQSRPSRNSL 201
gi|4502163|ref|NP_001638.1|       NCPKLS----- 189
gi|732003|sp|P39281|BLC_ECOLI     GS--------- 177
gi|127528|sp|P11590|MUP4_MOUSE    RCLKARE---- 178
```

图 10-3 对 5 个远缘的载脂蛋白进行多重比对（见图 10-2），结果来自 ClustalW。在第三步运用渐进序列比对建立多重比对。首先对齐两个最近的序列，然后根据在导向树中的位置将其他序列加进来。星号显示的是序列上 100% 的保守区域，冒号显示保守替换，点显示非保守替换

阶段1: 产生一系列的两两比对

```
Sequence format is Pearson
Sequence 1: gi|5803139|ref|NP_006735.1|          199 aa
Sequence 2: gi|6174963|sp|Q00724|RETB_MOUS       201 aa
Sequence 3: gi|132407|sp|P04916|RETB_RAT         201 aa
Sequence 4: gi|89271|pir||A39486                 201 aa
Sequence 5: gi|132403|sp|P18902|RETB_BOVIN       183 aa
```

```
Start of Pairwise alignments
Aligning...
Sequences (1:2) Aligned. Score:  84
Sequences (1:3) Aligned. Score:  84
Sequences (1:4) Aligned. Score:  91
Sequences (1:5) Aligned. Score:  92
Sequences (2:3) Aligned. Score:  99
Sequences (2:4) Aligned. Score:  86
Sequences (2:5) Aligned. Score:  85
Sequences (3:4) Aligned. Score:  85
Sequences (3:5) Aligned. Score:  84
Sequences (4:5) Aligned. Score:  96
```

阶段2: 计算距离矩阵产生辅助导向树

```
(
(
gi|5803139|ref|NP_006735.1|:0.04284,
(
gi|6174963|sp|Q00724|RETB_MOUS:0.00075,
gi|132407|sp|P04916|RETB_RAT:0.00423)
:0.10542)
:0.01900,
gi|89271|pir||A39486:0.01924,
gi|132403|sp|P18902|RETB_BOVIN:0.01902);
```

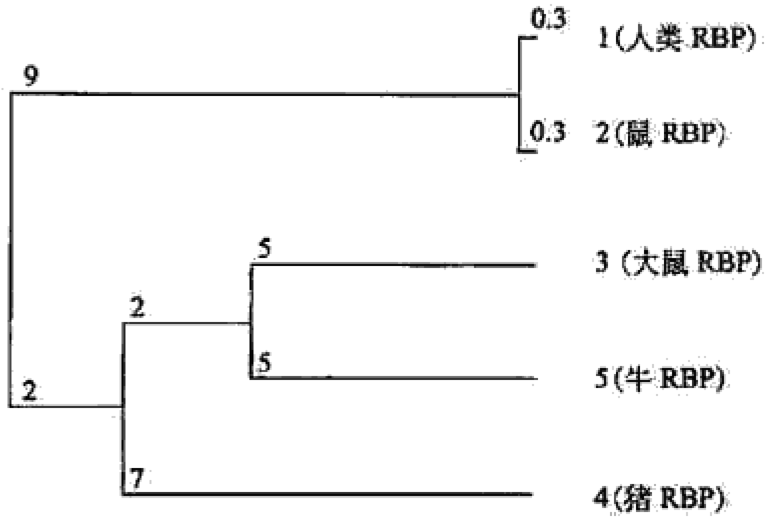


图 10-4 利用渐进比对算法，用 ClustalW 软件得到的多重比对。与远缘蛋白比较（图 10-2），此处两两比对的分值要高得多，距离（树枝长度）则小得多

Stage 3: progressively align the sequences following the branching order of the tree

```
CLUSTAL W (1.82) multiple sequence alignment

gi|89271|pir||A39486             MEWVWALVLLAALGSAQAERDCRVSSFRVKENFDKARFSGTWYAMAKKDP 50
gi|132403|sp|P18902|RETB_BOVIN   ------------------ERDCRVSSFRVKENFDKARFAGTWYAMAKKDP 32
gi|5803139|ref|NP_006735.1|      MKWVWALLLLAAW--AAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDP 48
gi|6174963|sp|Q00724|RETB_MOUS   MEWVWALVLLAALGGGSAERDCRVSSFRVKENFDKARFSGLWYAIAKKDP 50
gi|132407|sp|P04916|RETB_RAT     MEWVWALVLLAALGGGSAERDCRVSSFRVKENFDKARFSGLWYAIAKKDP 50
                                                   ********************:* ***:*****

gi|89271|pir||A39486             EGLFLQDNIVAEFSVDENGHMSATAKGRVRLLNNWDVCADMVGTFTDTED 100
gi|132403|sp|P18902|RETB_BOVIN   EGLFLQDNIVAEFSVDENGHMSATAKGRVRLLNNWDVCADMVGTFTDTED 82
gi|5803139|ref|NP_006735.1|      EGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTED 98
gi|6174963|sp|Q00724|RETB_MOUS   EGLFLQDNIIAEFSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTED 100
gi|132407|sp|P04916|RETB_RAT     EGLFLQDNIIAEFSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTED 100
                                 *********:*******.*:************.**:**************

gi|89271|pir||A39486             PAKFKMKYWGVASFLQKGNDDHWIIDTDYDTYAAQYSCRLQNLDGTCADS 150
gi|132403|sp|P18902|RETB_BOVIN   PAKFKMKYWGVASFLQKGNDDHWIIDTDYETFAVQYSCRLLNLDGTCADS 132
gi|5803139|ref|NP_006735.1|      PAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADS 148
gi|6174963|sp|Q00724|RETB_MOUS   PAKFKMKYWGVASFLQRGNDDHWIIDTDYDTFALQYSCRLQNLDGTCADS 150
gi|132407|sp|P04916|RETB_RAT     PAKFKMKYWGVASFLQRGNDDHWIIDTDYDTFALQYSCRLQNLDGTCADS 150
                                 ****************:*******:****:*:* ****** *********

gi|89271|pir||A39486             YSFVFARDPHGFSPEVQKIVRQRQEELCLARQYRLITHNGYCDGKSERNI 200
gi|132403|sp|P18902|RETB_BOVIN   YSFVFARDPSGFSPEVQKIVRQRQEELCLARQYRLIPHNGYCDGKSERNI 182
gi|5803139|ref|NP_006735.1|      YSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNL 198
gi|6174963|sp|Q00724|RETB_MOUS   YSFVFSRDPNGLSPETRRLVRQRQEELCLERQYRWIEHNGYCQSRPSRNS 200
gi|132407|sp|P04916|RETB_RAT     YSFVFSRDPNGLTPETRRLVRQRQEELCLERQYRWIEHNGYCQSRPSRNS 200
                                 *****:*** *:.**.:::********** **** * *****:.:..**

gi|89271|pir||A39486             L 201
gi|132403|sp|P18902|RETB_BOVIN   L 183
gi|5803139|ref|NP_006735.1|      L 199
gi|6174963|sp|Q00724|RETB_MOUS   L 201
gi|132407|sp|P04916|RETB_RAT     L 201
                                 *
```

图 10-5 5 个近邻视黄醇结合蛋白（retinol-binding protein）的多重比对。输出结果来自 ClustalW（用的是 Feng 和 Doolittle 的渐进比对算法）

(a)

	12345
人类RBP	GTWYA
鼠RBP	GLWYA
脂蛋白D	GRWYE
大肠杆菌BLC	GTWYE
鼠MUP4	GEWFS

(b)

概率	位置 1	2	3	4	5
p(G)	1.0				
p(T)		0.4			
p(L)		0.2			
p(R)		0.2			
p(E)		0.2			0.4
p(W)			1.0		
p(Y)				0.8	
p(F)				0.2	
p(A)					0.4
p(S)					0.2

(c) $p(GEWYE) = (1.0)(0.2)(1.0)(0.8)(0.4) = 0.064$

对数分值 $= \ln(1.0) + \ln(0.2) + \ln(1.0) + \ln(0.8) + \ln(0.4) = -2.75$

(d)

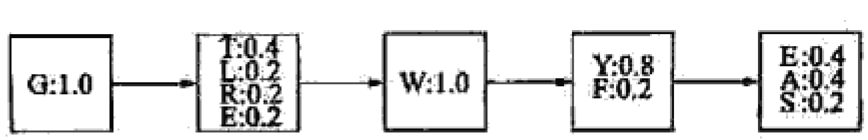


图 10-6 隐马模型描述在每个位点氨基酸残基出现的概率，这在概念上和 PSI-BLAST 所用的位置特异性打分矩阵是一致的。(a) 显示 5 个载脂蛋白 GXW 保守区域（图 10-3）的比对；(b) 计算每个氨基酸残基在各个比对位置的概率；(c) 通过这些概率，可以对每一个查询序列（如 GEWYE）给出一个分值，注意到真实的分值应该考虑空白和其他参数，这同样也是一种位置特异性的记分方案，例如同一个谷氨酸残基在位置 2 和位置 5 的出现概率是不同的；(d) 比对中与各个位置联系的氨基酸出现概率可以用状态转移表示

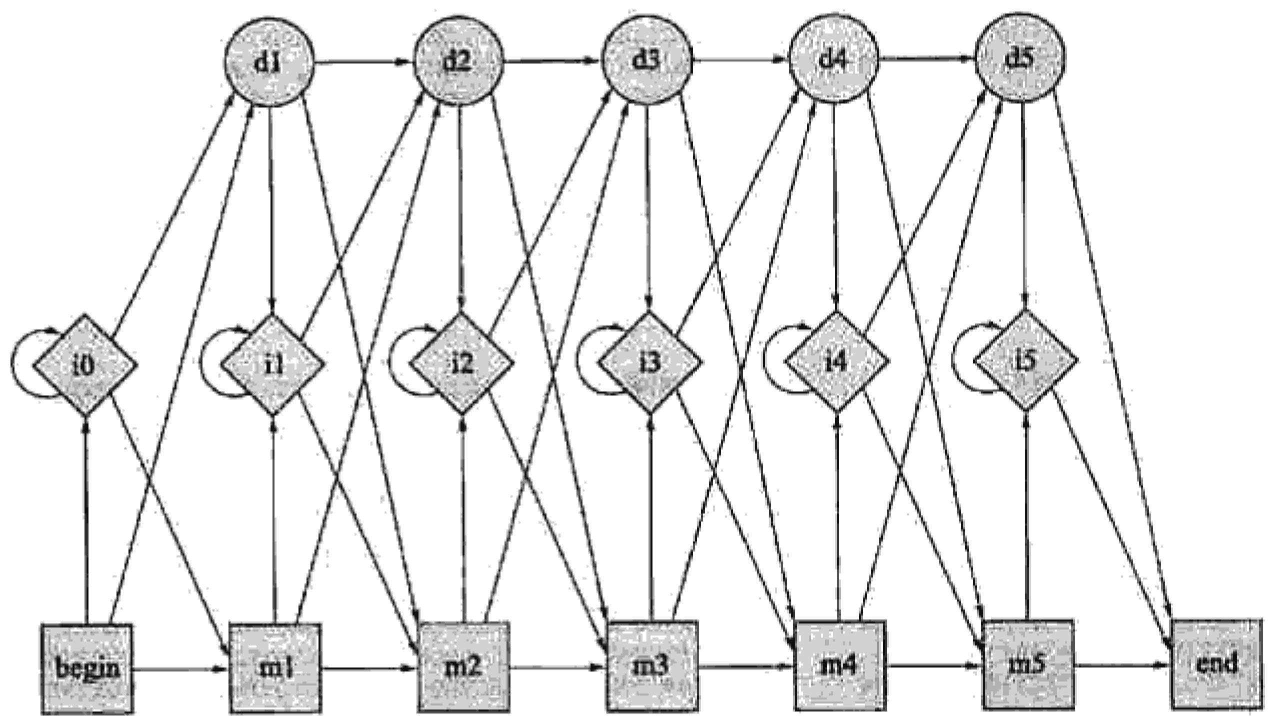


图 10-7 一个隐马模型的结构。HMM 由一系列状态组成，每一个状态与一定的概率联系，最底层是一列主要状态（从“begin”到 m1～m5，到“end”），这些状态对应多重比对中对齐的氨基酸残基，概率分布就是各残基的出现频率［图 10-6(d)］。第二行由插入状态组成（图 10-7，菱形图标 i1～i5）。这些状态模拟比对中的可变区域，允许在适当的地方进行插入。顶部第三行的圆形图标称为删除状态，它们对应了比对中的空白

HMMs 提供了一种搜索数据库中远缘同源的有力方法，对标准的 BLAST 搜索是一个很好的补充。HMMs 可以定义一个蛋白质或基因家族，现在已经实验了对基于 HMM 的数据库的搜索。网上基于 HMM 的三个最常用的数据库是 Pfam、SMART、TIGRFAMS（见下文介绍）。

HMMs 可以用 HMMER 程序产生。你可以用 hmmbuild 程序建立可以读人一个多序列比对的 profile HMM［图 10-8(a)］。接着 hmmcalibrate 程序将 profile HMM 和一组 5000 个随机序列匹配，将分值拟合为一个极值分布（extreme-value distribution，EVD)，计算 EVD 的各项参数，这对于评估数据库匹配的统计显著性是必需的［图 10-8(b)］。然后就可以用 hmmsearch 程序对数据库进行搜索。

当从 5 个高度分歧载脂蛋白的比对建立 profile HMM，对 GenBank 的非冗余数据库进行搜索，会得到很多匹配［图 10-9(a)］，包括 RBP4［图 10-9(b)］和细菌载脂蛋白［图 10-9(c)］。相比之下，用 5 个近邻的 RBP 直系同源建立的 profile HMM，查询结果仅仅是 RBPs［图 10-10(a)］，没有查寻到细菌的同源物。人类 RBP4 的比对有较高的分值和较低的 *E* 值。Profile HMM 对作为输入的多序列比对很敏感。

HMMER 的两个主要用途是利用一个 profile HMM 对数据库进行搜索（图 10-9、图 10-10），或者对一个 HMM 数据库查找一个序列。Pfam 就是这样一种数据库。

(a)

```
root@localhost hmmer]# hmmbuild lipocalins.hmm x.msf
hmmbuild - build a hidden Markov model from an alignment
HMMER 2.2g (August 2001)
Copyright (C) 1992-2001 HHMI/Washington University School of Medicine
Freely distributed under the GNU General Public License (GPL)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Alignment:        #1
Number of sequences: 5
Number of columns:  234

Determining effective sequence number   ... done. [4]
Weighting sequences heuristically       ... done.
Constructing model architecture         ... done.
Converting counts to probabilities      ... done.
Setting model name, etc.                ... done. [x]

Constructed a profile HMM (length 230)
Average score:    411.45 bits  <------------ 1
Minimum score:     353.73 bits
Maximum score:     460.63 bits
Std. deviation:   52.58 bits

Finalizing model configuration          ... done.
Saving model to file                    ... done.
//
```

(b)

```
root@localhost hmmer]# hmmcalibrate lipocalins.hmm
hmmcalibrate -- calibrate HMM search statistics
HMMER 2.2g (August 2001)
Copyright (C) 1992-2001 HHMI/Washington University School of Medicine
Freely distributed under the GNU General Public License (GPL)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
HMM file:              lipocalins.hmm
Length distribution mean: 325
Length distribution s.d.: 200
Number of samples:       5000
random seed:             1034351005
histogram(s) saved to:   [not saved]
POSIX threads:           2
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

HMM    : x
mu     : -123.894508
lambda :  0.179608
max    : -79.334000
//
```

图 10-8 HMMER 程序用一个多重比对作为输入建立 profile HMM。这个程序可以从 http://hmmer.wustl.edu 获得，在 linux 系统下运行。图 10-1～图 10-3 中的 5 个载脂蛋白对齐后输入 HMMER。(a) 用 hmmbuild 程序建立 profile HMM，(b) 用 5000 个序列的数据库进行校准。计算得到的 Mu 和 lambda 值可以对 HMM 进行统计评估。这个 profile HMM 的平均分值是 411bits [(a) 中的箭头 1]。当用 5 个近邻的视黄醇结合蛋白建立 profile HMM 时，平均分是 (600.89±24) bits

(a)

```
Scores for complete sequences (score includes all domains):
Sequence                              Description     Score    E-value  N
--------                              -----------     -----    -------  ---
gi|20888903|ref|XP_129259.1|          (XM_129259) ret  461.1   1.9e-133  1
gi|132407|sp|P04916|RETB_RAT          Plasma retinol-  458.0   1.7e-132  1
gi|20548126|ref|XP_005907.5|          (XM_005907) sim  454.9   1.4e-131  1
gi|5803139|ref|NP_006735.1|           (NM_006744) ret  454.6   1.7e-131  1
gi|20141667|sp|P02753|RETB_HUMAN      Plasma retinol-  451.1   1.9e-130  1
.
.
gi|16767588|ref|NP_463203.1|          (NC_003197) out  318.2   1.9e-90   1
```

(b)

```
gi|5803139|ref|NP_006735.1|: domain 1 of 1, from 1 to 195: score 454.6, E = 1.7e-131
                     *->mkwVMkLLLLaALagvfgaAErdAfsvgkCrvpsPPRGfrVkeNFDv
                        mkwV++LLLLaA  +   +aAErd      Crv+s    frVkeNFD+
  gi|5803139      1     MKWVWALLLLAA--W--AAAERD------CRVSS----FRVKENFDK 33

                     erylGtWYeIaKkDprFErGLllqdkItAeySleEhGsMsataeGrirVL
                     +r++GtWY++aKkDp  E GL+lqd+I+Ae+S++E+G+Msata+Gr+r+L
  gi|5803139     34 ARFSGTWYAMAKKDP--E-GLFLQDNIVAEFSVDETGQMSATAKGRVRLL 80

                     eNkelcADkvGTvtqiEGeasevfLtadPaklklKyaGvaSflqpGfddy
                     +N+++cAD+vGT+t++E          dPak+k+Ky+GvaSflq+G+dd+
  gi|5803139     81 NNWDVCADMVGTFTDTE----------DPAKFKMKYWGVASFLQKGNDDH 120

                     wIlaTDYdYnYAlvYSCtpiirLinEkDGkcadsyswilsRdPnGLspEt
                     wI++TDYd +YA++YSC    rL+n +DG+cadsys+++sRdPnGL+pE+
  gi|5803139    121 WIVDTDYD-TYAVQYSC----RLLN-LDGTCADSYSFVFSRDPNGLPPEA 164

                     keklrkvlteegidvkkqyiwitqnnyCpkrse<-*
                     ++++r+  ++e++++++qy++i++n+yC++rse
  gi|5803139    165 QKIVRQ--RQEELCLARQYRLIVHNGYCDGRSE    195
```

(c)

```
gi|16767588|ref|NP_463203.1|: domain 1 of 1, from 1 to 177: score 318.2, E = 1.9e-90
                     *->mkwVMkLLLLaALagvfgaAErdAfsvgkCrvpsPPRGfrVkeNFDv
                            M+LL+ +A       a ++ Af+v++C++p+PP+G++V++NFD+
  gi|1676758      1     ----MRLLPVVA------AVTA-AFLVVACSSPTPPKGVTVVNNFDA 36

                     erylGtWYeIaKkDprFErGLllqdkItAeySleEhGsMsataeGrirVL
                     +rylGtWYeIa+ D+rFErGL   + +tA+ySl++         +G+i+V+
  gi|1676758     37 KRYLGTWYEIARLDHRFERGL---EQVTATYSLRD--------DGGINVI 75

                     eNkelcADkvGTvtqiEGeasevfLtadPaklklKyaGvaSflqpGfddy
                      Nk++++D+   +++ +EG+a   ++t+ P +++lK+    Sf++p++++y
  gi|1676758     76 -NKGYNPDR-EMWQKTEGKA---YFTGSPNRAALKV----SFFGPFYGGY 116

                     wIlaTDYdYnYAlvYSCtpiirLinEkDGkcadsyswilsRdPnGLspEt
                     +++a+D++Y++Alv  C+p      + +D      y+wilsR+P+ Ls E+
  gi|1676758    117 NVIALDREYRHALV--CGP-----D-RD------YLWILSRTPT-LSEEI 151

                     keklrkvlteegidvkkqyiwitqnnyCpkrse<-*
                     k+++++v+t+eg+dv k  iw++q      +  +
  gi|1676758    152 KQQMLAVATREGFDVNK-LIWVKQ----SG--S    177
```

图 10-9 (a) 用 5 个近邻的载脂蛋白训练 HMM 时，HMMER 的输出包括了一组载脂蛋白，其中有来自细菌的成员。下面显示的是 HMM 和 (a) 人类 RBP4 与 (b) 细菌载脂蛋白的比对

(a)

```
Scores for complete sequences (score includes all domains):
Sequence                                   Description         Score    E-value  N
--------                                   -----------         -----    ------- ---
gi|3041715|sp|P27485|RETB_PIG              Plasma retinol-     614.2   1.6e-179   1
gi|89271|pir||A39486                       plasma retinol-     613.9   1.9e-179   1
gi|20888903|ref|XP_129259.1|               (XM_129259) ret     608.8   6.8e-178   1
gi|132407|sp|P04916|RETB_RAT               Plasma retinol-     608.0   1.1e-177   1
gi|20548126|ref|XP_005907.5|               (XM_005907) sim     607.3   1.9e-177   1
gi|20141667|sp|P02753|RETB_HUMAN           Plasma retinol-     605.3   7.2e-177   1
gi|5803139|ref|NP_006735.1|                (NM_006744) ret     600.2   2.6e-175   1
```

(b)

```
gi|5803139|ref|NP_006735.1|: domain 1 of 1, from 1 to 199: score 600.2, E = 2.6e-175
                   *->meWvWaLvLLaalGgasaERDCRvssFRvKEnFDKARFsGtWYAiAK
                      m+WvWaL+LLaa+  a+aERDCRvssFRvKEnFDKARFsGtWYA+AK
  gi|5803139     1    MKWVWALLLLAAW--AAAERDCRVSSFRVKENFDKARFSGTWYAMAK 45

                   KDPEGLFLqDnivAEFsvDEkGhmsAtAKGRvRLLnnWdvCADmvGtFtD
                   KDPEGLFLqDnivAEFsvDE+G+msAtAKGRvRLLnnWdvCADmvGtFtD
  gi|5803139    46 KDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTD 95

                   tEDPAKFKmKYWGvAsFLqkGnDDHWiiDtDYdtfAvqYsCRLlnLDGtC
                   tEDPAKFKmKYWGvAsFLqkGnDDHWi+DtDYdt+AvqYsCRLlnLDGtC
  gi|5803139    96 TEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTC 145

                   ADsYsFvFsRDPnGlsPEvqkivRqRqEELCLaRqYRlieHnGYCdgrse
                   ADsYsFvFsRDPnGl+PE+qkivRqRqEELCLaRqYRli+HnGYCdgrse
  gi|5803139   146 ADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSE 195

                   RniL<-*
                   Rn+L
  gi|5803139   196 RNLL    199
```

图 10-10 (a) 用近邻的 5 个 RBP 多重比对训练 HMMER 程序，数据库搜索的结果主要由 RBP 的直系同源组成；(b) 注意这里的 E 值比图 10-9 中 HMM 的低，人类 RBP 与 HMM 的比对分值较高，然而这次搜索没有找到细菌等的远缘蛋白

10.2 两种多序列比对的程序

我们将用介绍研究蛋白质家族的网上资源的方式来研究两种主要的多重序列比对程序。

① 现在可以获得很多的多重比对数据库，可以用文本（关键字）或者任意序列进行搜索，查询的可以是已知序列（如 RBP）或新的序列（如你新鉴定的一种载脂蛋白）。在某些数据库里，你提供的查询序列将加入到一组预先计算好的蛋白家族多重比对中。

② 还可以用手动方式对输入的一组蛋白质或核酸序列产生多重比对。在这种情况下，比对是根据输入序列得到的，没有对数据库进行搜索。

10.2.1 多序列比对的数据库

很多组织建立了大型的蛋白质家族和结构域的数据库（表 10-1）。下面对这些资源的特点，相互之间的联系、区别作一介绍。

10.2.1.1 Pfam：基于 profile HMM 的蛋白家族数据库

Pfam 是综合的蛋白质家族数据库之一（Bateman 等，2000，2002），它同时收集了序列多重比对和蛋白家族的 profile HMMs，可以用文本（关键字、蛋白质名等）或序列数据进行搜索。Pfam 属于 Wellcome Trust Sanger Institue（Hinxton，United Kindom），在美国、

表 10-1 多序列比对数据库

数据库名称	描 述	网 站 地 址
BLOCKS	类隐与模型库，无空隙	http://www.blocks.fhcrc.org/
CDD	保守结构域数据库	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
Interpro	联合了 PROSITE，PRINTS，ProDom Pfam，SMART，TIGRfam 的资源	http://www.ebi.ac.uk/interpro/index.html
iProClass database	来自 PIR	http://pir.georgetown.edu/iproclass
MetaFAM	联合了 6 个数据库的蛋白质数据	http://metafam.ahc.umn.edu/
Pfam	隐马模型库	http://Pfam.wustl.edu/http://www.sanger.ac.uk/Software/Pfam/
PIR-ALN	比对数据库	
PRINTS	SwissProt/TrEMBL 的蛋白指纹	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
ProDom	利用 PSI-BLAST 对 Swiss Prot 蛋白质数据聚类	http://prodes.toulouse.inra.fr/prodom/doc/prodom.html
PROSITE	蛋白质模体字典	http://www.expasy.ch/prosite/
SMART	简单分子构架研究工具	http://smart.embl-heidelberg.de/
TIGRFAMs	蛋白家族的隐马模型库	http://tigrblast.tigr.org/web-hmm/

瑞典、法国有镜像。

Pfam 由两个数据库组成。Pfam-A 是手工编辑、多重比对形式的蛋白质家族集合。对于每一个家族，Pfam 提供了 4 种特征：注释、种子比对、profile HMM、完全比对。完全比对可能很大，Pfam 前 20 个家族的完全比对都含有超过 2500 个序列。种子比对含有较少数量的代表序列。除了由专家手工编辑的 Pfam-A 外，Pfam-B 是从 ProDom 数据库自动产生（见下文）的。Pfam-B 的数据质量、注释的完全程度都不如 Pfam-A，但是 Pfam-B 能使整个数据库更加综合，可以作为一个有用的补充。

> Pfam 有 4 个 URL：http://Pfam.wustl.edu/（U.S.），http://www.sanger.ac.uk/Software/Pfam/（U.K.），http://www.cgb.ki.se/Pfam（Sweden），http://pfam.jouy.inra.fr（France）。其 9.0 版（2003，5 月）包含了 5724 个蛋白质家族。目前，SwissProt 40 和 TrEMBL 中超过 69% 的蛋白质至少与 Pfam-A 中一个家族匹配。
> 你可以用一个 DNA 序列对 Pfam 进行查询：http://www.sanger.ac.uk/Software/Pfam/dnasearch.shtml。
> SMART（http://smart.embl-heidelberg.de/）目前拥有大约 650 个 profile HMM。

通过查询载脂蛋白，我们可以看到 Pfam 的主要特征。有 3 种方法访问数据库：浏览每个家族，搜索蛋白质序列（利用获取码或者序列），利用文本搜索。在主页面上，选择文本方式搜索，输入"lipocalin"（图 10-11）。输出结果有 4 部分（图 10-12）。最上面是注释，包含了访问其他数据库中与载脂蛋白相关记录的链接。接下来是多种格式的多序列比对文件，部分输出结果显示于图 10-13。另一个多功能的输出形式是 JalView。选择这个选项后，点击 JalView 按钮（图 10-14），可以在 applet 窗口中查看序列比对并进行分析和保存（图 10-15）。这个 applet 还能显示比对家族的主成分分析（PCA）结果（图 10-16）。PCA 是一种将数据降维的技术（图 7-16）。在这里，根据一种距离标度将多序列比对中的每一个蛋白质表示为空间中的一个点，因此很容易鉴别不属于同一类的蛋白。类似的信息可以用树的形式在 applet 中表示出来。

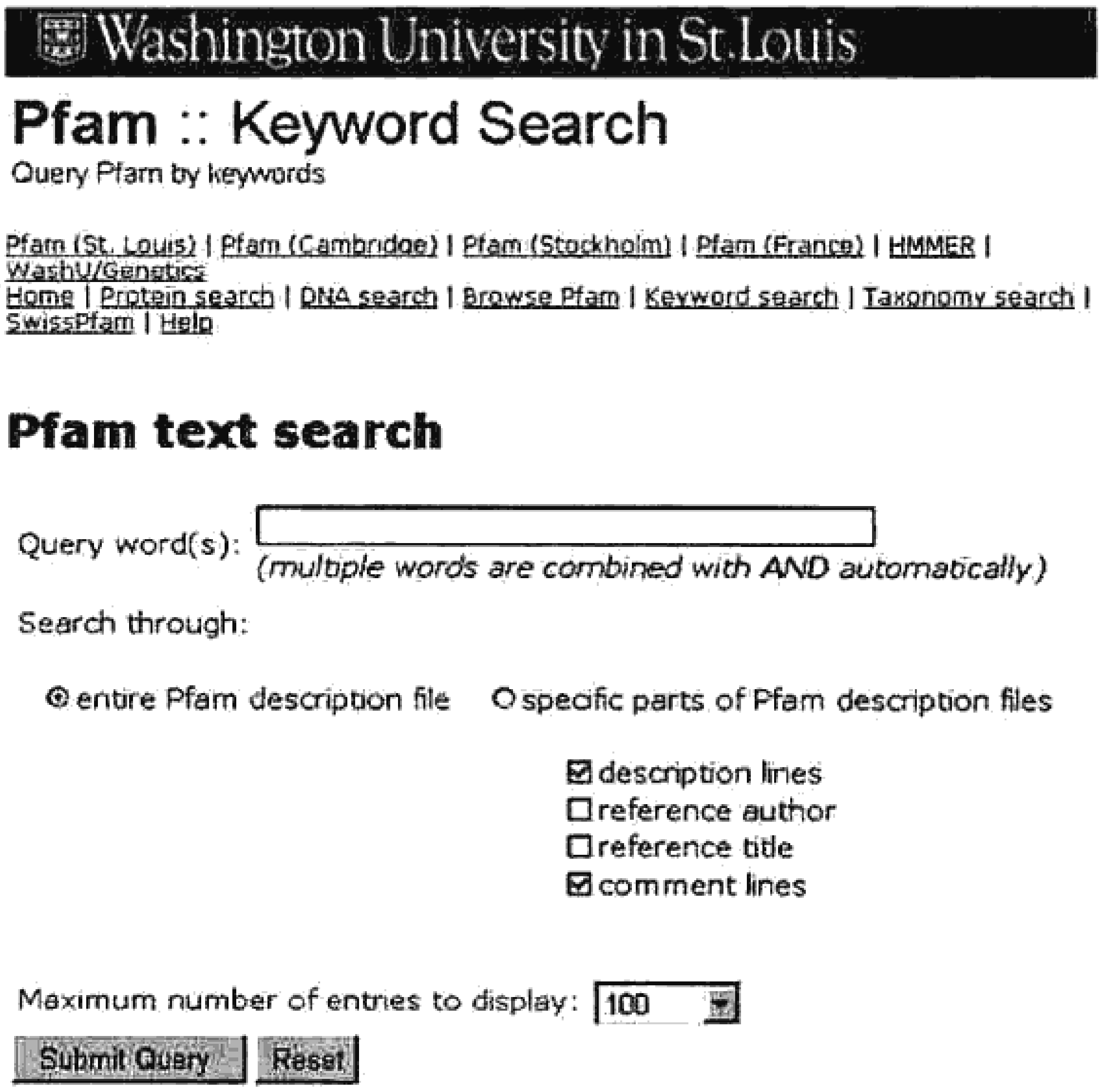


图 10-11　Pfam 数据库允许使用文本形式的搜索（见 http://pfam.wustl.edu/）

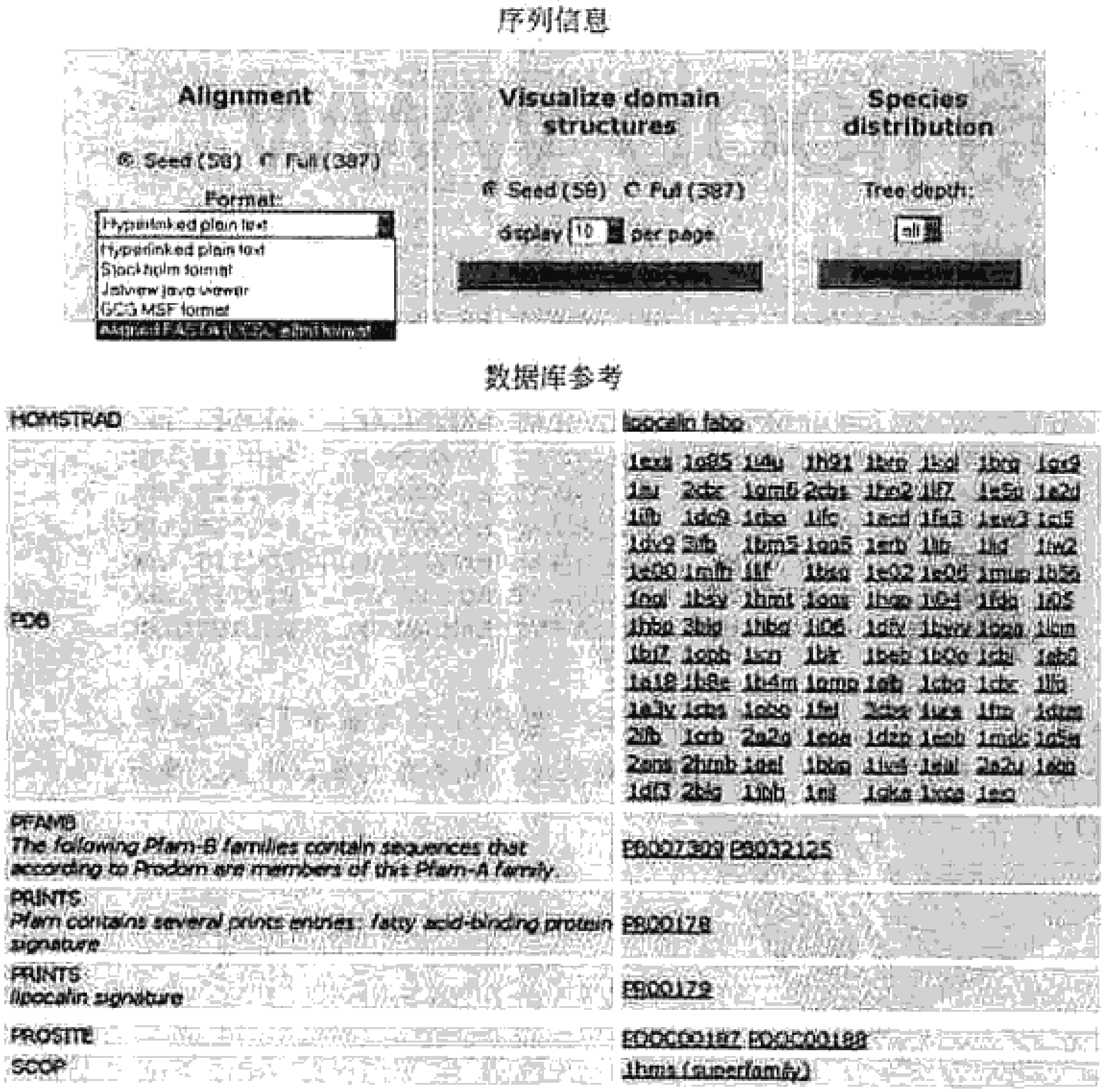


图 10-12　在 Pfam 用关键字“lopocalin”进行搜索，结果得到了载脂蛋白/胞浆脂肪酸结合蛋白家族（lipocalin/cytosolic fatty acid-binding protein family）。部分输出结果如图所示，获取码为 PF00061，作者为 Sean Eddy（即 Pfam 的作者）。结果提供了访问其他数据库（关于蛋白家族、模体等）的链接。Pfam 的多重比对可以用很多种格式显示。与查看 profile HMM 类似，可以查看结构域的结构

```
     VEGP_HUMAN_30-171   DVSGTWY.LK AMTV.DREFP ..EMNLESVT PMTLTTLE.G GNLEAK.VTM
1    VEG1_RAT_29-172     DVSGTWY.LK AAAW.DKEIP DKKFGSVSVT PMKIKTLE.G GNLQVK.FTV
-->  LALP_MACEU_28-171   PSEGTYY.VQ VIAV.DKEFP .EDEIPRDIS PLTITYLN.N GKMEAK.FTV
     PGHD_HUMAN_38-186   KFLGRWF.SA GLASNSSWLR .EKKAALSMC KSVVAPAT.D GGLNLT.STF
     LIPO_BUFMA_32-179   KILGKWY.GI GLASNSNWFQ .SKKQQLKMC TTVITPTA.D GNLDVV.ATF
     NGAL_MOUSE_46-197   QFRGRWY.VV GLAG.NAVQK ..KTEGSFTM YSTIYELQ.E NNSYNV.TSI
     NGAL_HUMAN_46-195   QFQGKWY.VV GLAG.NAILR .EDKDPQKMY AT.IYELK.E DKSYNV.TSV
     AMBP_PLEPL_41-189   RFVGTWH.DV ALTSSCPHMQ ..RNRADAAI GKLVLEKDTG NKLKVT.RTR
     AMBP_HUMAN_39-188   RIYGKWY.NL AIGSTCPWLK .KIMDRMTVS TLVLGEGATE AEISMT.STR
     CO8G_HUMAN_46-188   QFAGTWL.LV AVGSACRFLQ .EQGHRAEAT TLHVAPQG.. TAMAVS.TFR
     ERBP_RAT_32-176     KFLGFWY.EI AFASKMGTPG ..LAHKEEKM GAMVVELK.E NLLALT.TTY
     EFA8_CHICK_29-173   EVAGKWY.IV ALASNTDFFL .REKGKMKMV MARISFLG.E DELEVS.YAA
     OLFA_RANPI_30-174   KVTGVWY.GI AAASNCKQFL QMKSDNMPAP VNIYSLNN.. GHMKSS.TSF
     PAEP_HUMAN_32-178   KLAGTWH.SM AMATNNISLM ATLKAPLRVH ITSLLPTP.E DNLEIV.LHR
     LACB_PIG_32-176     KVAGTWH.TV AMAVSDVSLL DAKSSPLKAY VEGLKPTP.E GDLEIL.LQK
     LACB_BOVIN_30-176   KVAGTWY.SL AMAASDISLL DAQSAPLRVY VEELKPTP.E GDLEIL.LQK
     LACA_CANFA_14-159   KVAGTWH.SM AMAASDISLL DSETAPLRVY IQELRPTP.Q DNLEIV.LRK
     LACB_EQUAS_14-160   EVAGKWH.SV AMAASDISLL DSEEAPLRVY IEKLRPTP.E DNLEII.LRE
     LACA_EQUAS_14-161   EVAGRWH.SV AMVASDISLL DSESAPLRVY VEELRPTP.E GNLEII.LRE
     LACB_MACGI_14-153   KFVGSWY.LR EAAK...TME .FSIPLFDMD IKEVNLTP.E GNLELV.LLE
     PBAS_RAT_27-170     KIEGNWR.TV YLAASSVEKI .NEGSPLRTY FRRIECGK.R CNRINL.YFY
     OBP_RAT_27-170      EVNGDWR.TL YIVADNVEKV .AEGGSLRAY FQHMECGDEC QELKII.FNV
     APHR_CRICR_21-165   ELQGKWY.TI VIAADNLEKI .EEGGPLRFY FRHIDCYKNC SEMEIT.FYV
     OBP_BOVIN_12-156    ELSGPWR.TV YIGSTNPEKI .QENGPFRTY FRELVFDDEK GTVDFY.FSV
     MUP_RAT_33-176      KLNGDWF.SI VVASNKREKI .EENGSMRVF MQHIDVLE.. NSLGFK.FRI
     MUP3_MOUSE_36-179   QISGYWF.SI AEASYEREKI .EEHGSMRAF VENITVLE.. NSLVFK.FHL
     MUP1_MOUSE_32-175   KINGEWH.TI ILASDKREKI .EDNGNFRLF LEQIHVLE.. NSLVLK.FHT
     ESP4_LACVV_33-167   KTVGKWH.PI GMASKLPEVP ..EYEQKISP MDHMVELT.D GDMKLT.ANY
2    |A1AG_RAT_39-183    WLSDKWF.YM GAAFRDPVFK .QAVQTIQTE YFYLTPNLIN DTIELR.EFQ
-->  |A1AH_MOUSE_39-184  WLSDKWF.FI GAAVLNPDYR .QEIQKTQMV FFNLTPNLIN DTMELR.EYH
     |A1AG_RABIT_38-183  QLSHKWF.FT ASAFRNPKYK .QLVQHTQAA FFYFTAIKEE DTLLLR.EYI
     A1AG_HUMAN_38-183   QITGKWF.YI ASAFRNEEYN .KSVQEIQAT FFYFTPNKTE DTIFLR.EYQ
     RETB_XENLA_39-194   RYAGVWY.AV AKKDPEGLFL .LDNIAANFK IEDNGKTT.A TAKGRV.RIL
     RETB_BOVIN_19-174   RFAGTWY.AM AKKDPEGLFL .QDNIVAEFS VDENGHMS.A TAKGRV.RLL
     RET1_ONCMY_18-173   RYTGRWY.AV AKKDPVGLFL .LDNVVAQFS VDESGKVT.A TAHGRV.IIL
     PURP_CHICK_39-196   RYAGKWY.AL AKKDPEGLFL .QDNISAEYT VEEDGTMT.A SSKGRV.KLF
     CRC1_HOMGA_30-168   SYAGVWY.QF ALTN.NPYQL IEKCVRNEYS FDGKQFVI.. KSTGIA.YDG
     CRA2_HOMGA_25-170   RYAGRWY.QT HIIE.NAYQP ..VTRCIHSN YEYSTNDY.. GFKVTT.AGF
     ICYA_MANSE_22-174   AFAGAWH.EI AKLPLENENQ .GKCTIAEYK YD.GKKAS.V YNSFVS.NGV
     BBP_PIEBR_36-184    NYHGKWW.EV AKYPNSVEKY .GKCGWAEYT PE.GKSVK.V SNYHVI.HGK
     APOD_HUMAN_41-185   KYLGRWY.EI EKIPTTFENG ..RCIQANYS LMENGKIK.V LNQELR.ADG
     RET3_MOUSE_2-136    NFAGTWK.MR SSENFDELLK ..ALGVNAML RKVAVAAASK PHVEIR.QDG
     MYP2_BOVIN_3-131    KFLGTWK.LV SSENFDEYMK ..ALGVGLAT RKLGNLAK.. PRVIIS.KKG
     FABA_HUMAN_3-131    AFVGTWK.LV SSENFDDYMK ..EVGVGFAT RKVAGMAK.. PNMIIS.VNG
     FABH_BOVIN_3-131    AFVGTWK.LV DSKNFDDYMK ..SLGVGFAT RQVGNMTK.. PTTIIE.VNG
     FABL_GINCI_3-131    AFLGSWK.LQ KSHNFDEYMK ..NLDVSLAQ RKVATTVK.. PKTIIS.LDG
     FABE_HUMAN_6-134    QLEGRWR.LV DSKGFDEYMK ..ELGVGIAL RKMGAMAK.. PDCIIT.CDG
     FABP_SCHMA_3-132    SFLGKWK.LS ESHNFDAVMS ..KLGVSWAT RQIGNTVT.. PTVTFT.MDG
     FABP_ECHGR_3-129    AFLGTWK.ME KSEGFDKIME ..RLGVDFVT RKMGNLVK.. PNLIVTDLGG
```

图 10-13 Pfam 比对可以用 GCG MSF 的格式查看（多重比对形式）。这里显示了部分结果。GXW 模体非常保守，整个比对中只有 LALP_MACEU（箭头 1）和来自大鼠、小鼠、兔子的 α1-酸性糖蛋白（箭头 2）不同

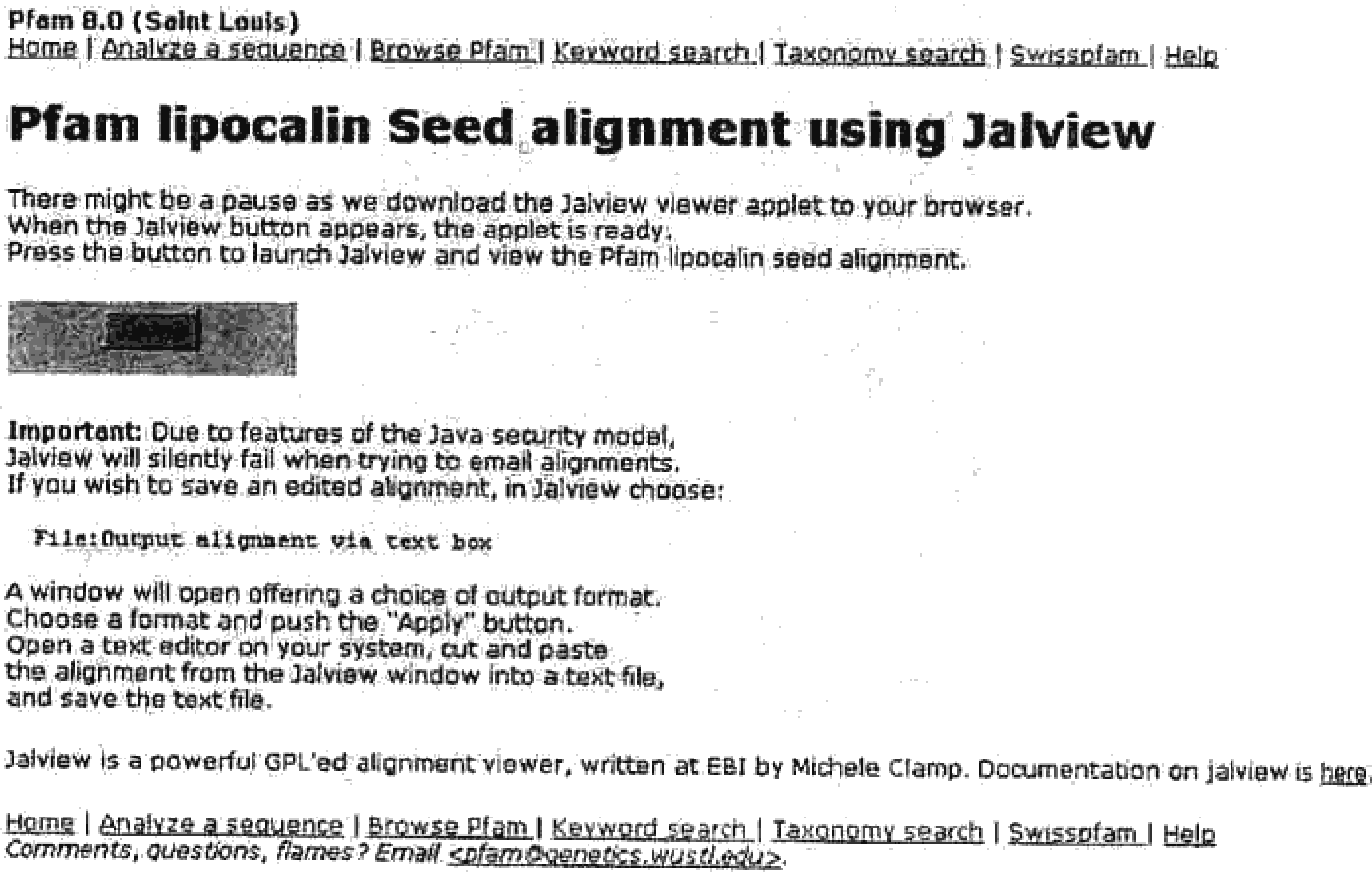


图 10-14 当出现 JalView 的方框时，点击后可以得到一个用 JalView 查看器显示的 Pfam 比对

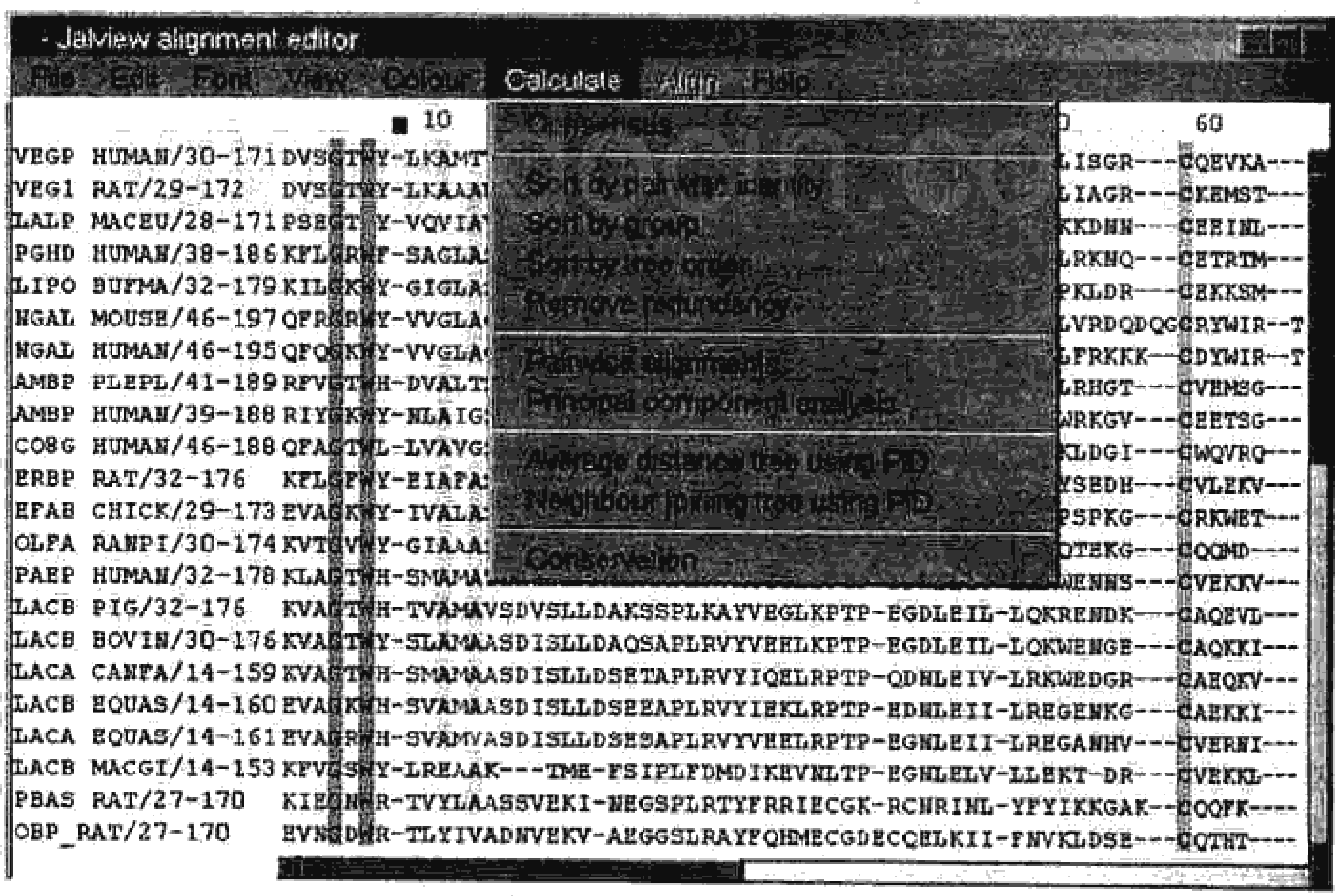


图 10-15 Pfam 的 Jalview 可显示 Pfam 中任意蛋白家族的多重比对。对齐的残基可用不同的颜色显示。同一个家族中不同蛋白质之间的关系可以用很多算法来分析，如主成分分析(PCA)(图 10-16) 和聚类树（图 10-17）等

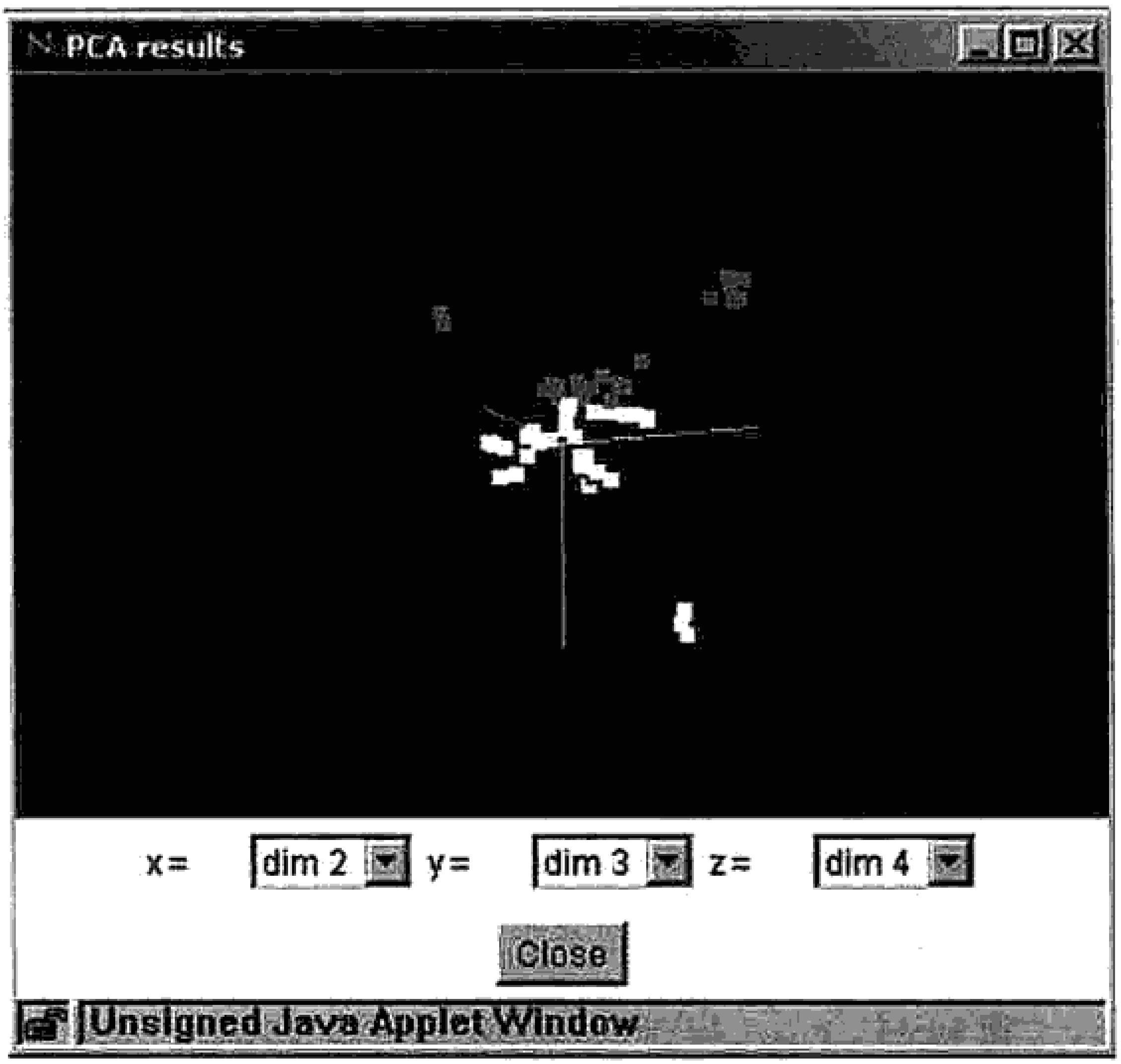


图 10-16 Pfam 的一个家族分析。用 JalView 显示 PCA 分析结果。家族中的每一个蛋白质表示为空间的一个点，坐标轴反映蛋白质之间的距离。可以用鼠标控制坐标轴旋转。很明显有几个载脂蛋白离群很远。选定任一蛋白，在 JalView 的多重比对中的相应序列就会高亮显示

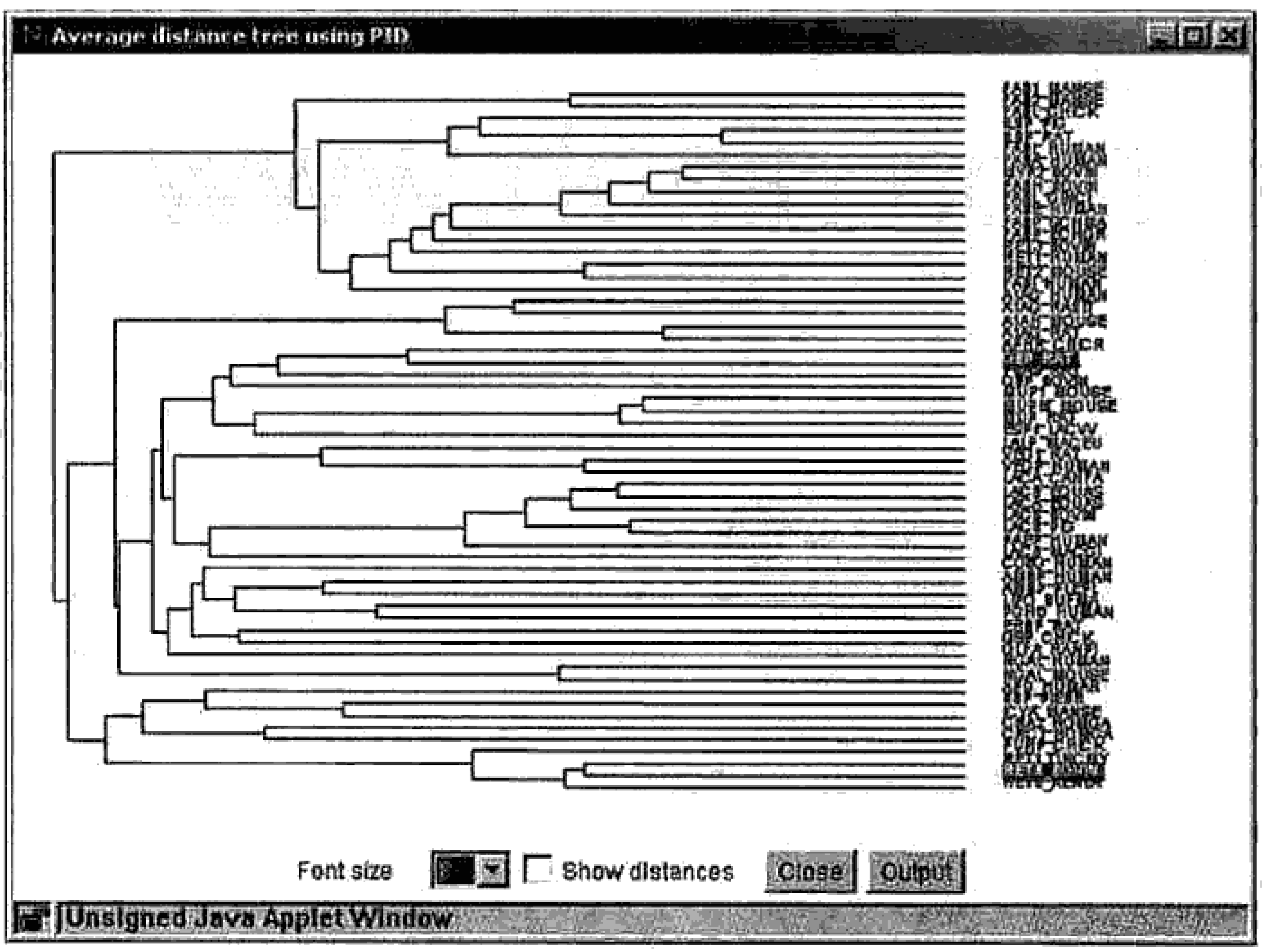


图 10-17 以树的形式显示的 Pfam 蛋白家族，这是 58 个载脂蛋白的平均距离树

10.2.1.2 SMART

简单分子构架研究工具（the simple molecular architecture research tool，SMART）数据库中的蛋白质涉及胞内信号转导，胞外结构域和染色质功能等方面（Schultz 等，1998；Ponting 等，1999）。像 Pfam 一样，SMART 使用了基于 profile HMMs 的 HMMER 软件。SMART 数据库也可以利用序列或关键词搜索（图 10-18）。搜索人类 RBP4 可以得到关于载脂蛋白结构域组织、组成的描述（图 10-19）。在 SMART 搜索得到的结构域具有更详尽的注释，包含了功能类型、三维结构、分类等信息。搜索酵母基因组所编码的蛋白，结果显示 6.7%的基因具有一个或多个信号结构域，比 SwissProt 和 Pfam 检测到的要多（Schultz 等，1998）。

> CDD 可以从 http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml 或从 BLAST 页面 http://www.ncbi.nlm.nih.gov/BLAST 访问。DART 可通过 http://www.ncbi.nlm.nih.gov/Structure/lexington/html/overview.html 访问。

10.2.1.3 保守结构域数据库（conserved domain database）

保守结构域数据库是 NCBI 的一个工具，允许基于序列或文本方式对 Pfam 和 SMART 进行查询。CDD 通过比较目标序列和一组位置特异性打分矩阵进行 RPS-BLAST。RPS-BLAST 和 PSI-BLAST 有联系也有区别，它对预先选定比对产生的模式进行搜索。使用 CDD 的目的是确定目标序列中的保守结构域。

以 HIV-1 pol 作为查询用 CDD 搜索的结果如图 10-20 所示，同时提供了分数和 *E* 值。对于每一个 PSSM 匹配（如鉴定出的每一个模体），查询序列将加入到那个家族的比对中。同样，如图 10-20 所示，还可以在比对中加入附加的序列，包括同一个家族中近邻的和远缘成员。

> 目前 BLOCKS 数据库（http://www.blocks.fhcrc.org/）拥有来自 2680 个类群的 11853 个 blocks。

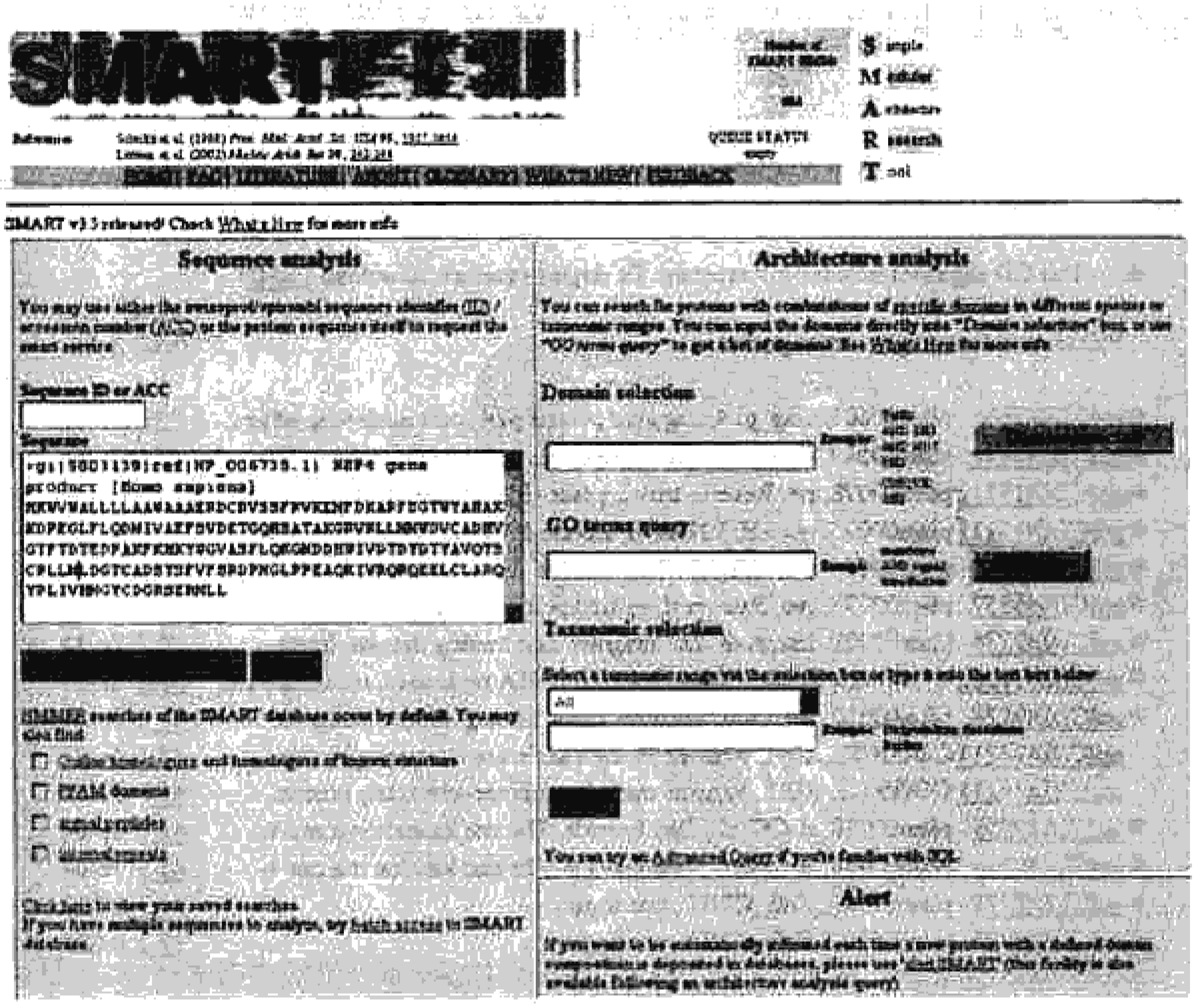


图 10-18 SMART 数据库（http://smart.embl-heidelberg.de/）允许搜索蛋白质序列的保守结构域。可以通过关键字（底部），结构域分析（右）或序列查询（上左，在查询框中粘贴 FASTA 格式的人类 RBP4 蛋白序列）进行搜索

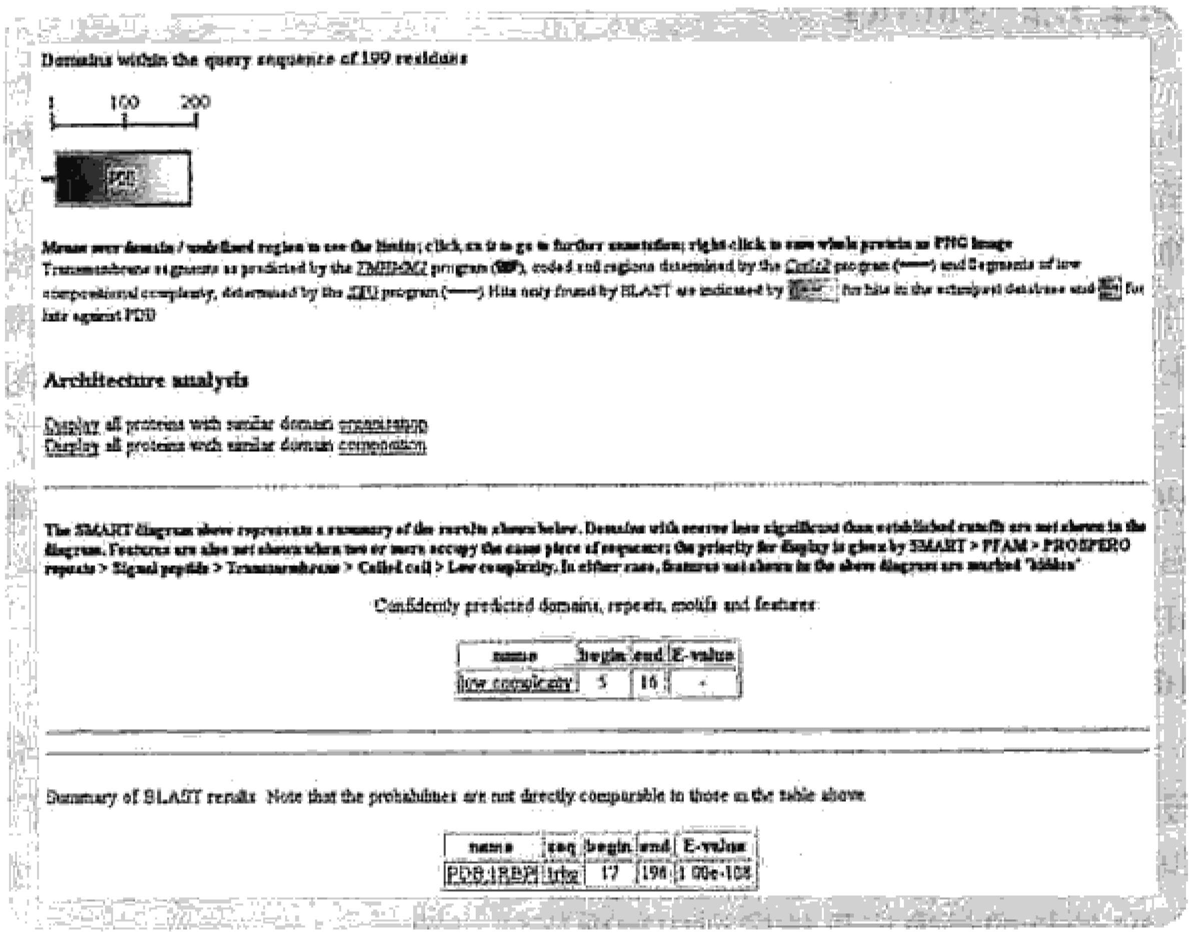


图 10-19 一次用人类RBP4作为查询的SMART搜索的结果。结果中有一个PDB中RBP记录的链接。结果鉴定出一个低复杂性的区域，对应于分泌蛋白特征的信号模体

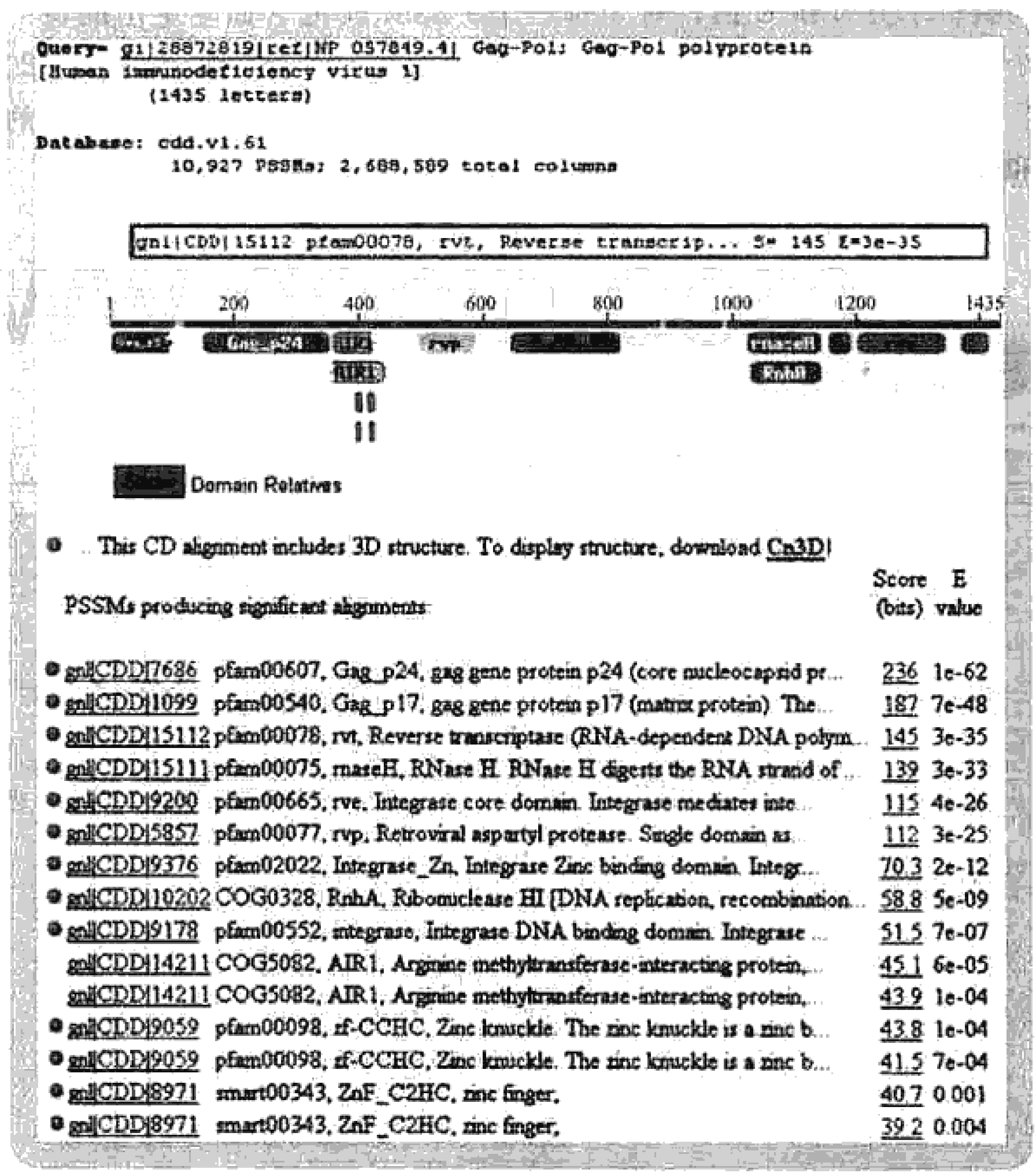


图 10-20 使用RPS-BLAST对CDD搜索得到的部分结果，查询目标是HIV-1 pol (NP _ 057849)和匹配都属于Pfam的蛋白家族

CDD 也可以通过在 NCBI 的 DART（domain architecture retrieve tool）中输入查询序列进行搜索。在 DART 中查询 HIV-1 pol 得到的结果如图 10-21。

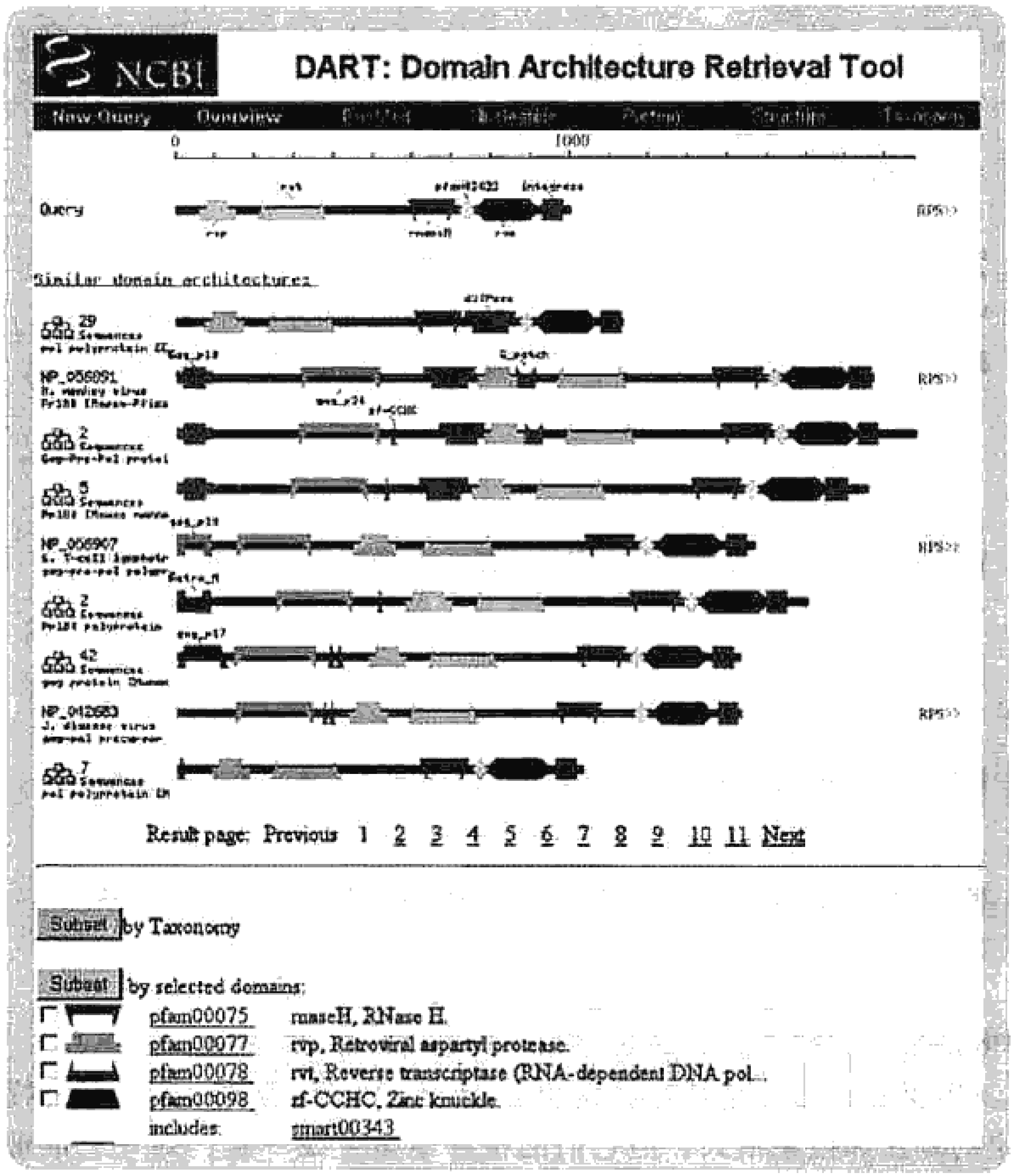


图 10-21 DART 是 NCBI 的一个工具，提供 CDD 搜索。可以通过 NCBI 结构页面的 CDD 站点进入 DART

10.2.1.4 BLOCKS

BLOCKS 数据库只包含对应于蛋白质高度保守区的无空白多重比对（Henikoff 等，1998，1999）。整合了另外 5 个数据库（PROSITE、Prints、ProDom、Domo、Pfam）的 Blocks＋已经建立起来。

31 版 PRINTS 拥有 1550 个数据库记录，总共覆盖了 9531 个模体。

10.2.1.5 PRINTS

PRINTS 数据库包含了定义于 SwissProt/TrEMBL 中的蛋白质家族的蛋白“指纹”。PRINTS 输出结果中有一个超链接是 CINEMA（the colour interactive editor for multiple alignment）编辑器（Parry Smith 等，1998）。这是一个与比对分析软件整合的 Java applet 程序。

10.2.1.6 PROSITE

PROSITE 数据库提供了对蛋白质结构域和蛋白家族的描述（Hofmann 等，1999；Sigrist 等，2002）。主页上包括了基于关键

最新版本（2003 年 6 月）的 PROSITE 包含了对超过 1000 个家族的描述。可以从通过面的地址访问 http://www.exasy.org/prosite/。

词的搜索（图 10-22）。对“lipocalin”的搜索会得到一个指向描述该家族的文件链接，我们曾经在图 8-8 遇到过这样的文件。

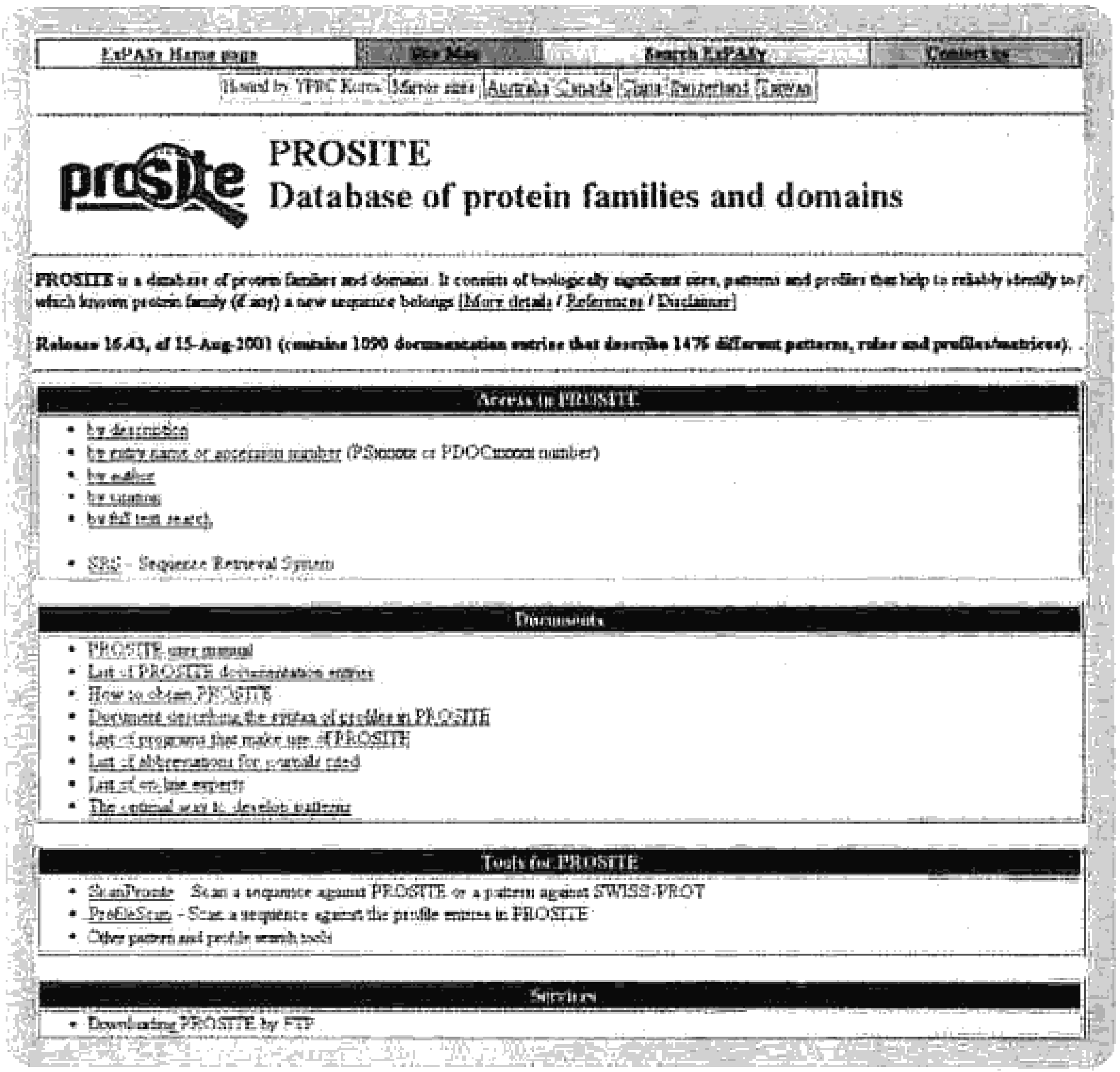


图 10-22 PROSITE 蛋白家族数据库（http://www.expasy.ch/prosite/）是蛋白质分析专家系统（the expert protein analysis system，ExPASy）的一部分。PROSITE 是一个关于蛋白质家族、结构域和类型的数据库，由 Amos Bairoch 建立，目前拥有超过 1000 个家族

10.2.1.7 综合的多重比对资源：InterPro，MetaFam，iProClass

InterPro 位于 http://www.ebi.ac.uk/interpro/。2002 年 9 月“Briefings in Bioinformatics”杂志主要介绍了 InterPro 和它的成员数据库。目前 85% 的 TrEMBL 蛋白序列在 InterPro 中有多于 1 条的匹配。
MetaFam 位于 http://metafam.abc.umn.edu/。

你可以通过 http://pir.georgetown.edu/iproclass/ 访问 iProClass。

多序列比对数据库的一个主题是虽然每个数据库都有自己特定的算法和搜索格式，但它们能很好地整合在一起。另一个很重要的思想是各个数据库（如 Pfam 和 PROSITE）都采用了特定的方法对蛋白质进行分类分析，有些数据库采用 HMM，有些主要考虑蛋白质结构域，另外一些则针对更小的模体。综合这些资源，你可以并行使用多种算法对某个蛋白质的一些特征进行深入研究。

目前至少发展了 3 个综合大多数多重比对数据库的资源。InterPro 数据库将 PROSITE、PRINTS、ProDom、Pfam、TIGRFAMs 和 BLOCKS 组织成交叉参考（表 10-2）（Mulder 等，2002，2003）。这项工程联合了 EBI、Wellcome Trust Sanger Institution 等 8 个中心（Apweiler 等，2001）。

另一个综合的资源是 MetaFam（Shoop 等，2001；Silverstein 等，2001a，2001b）。MetaFam 统一了来自 10 个数据库的比对（Blocks+、DOMO、Pfam、PIR-ALN、PRINTS、PROSITE、ProDom、PROTOMAP、SBASE、SYSTERS）（Silverstern，2001b）。

iProClass 将 PIR 和 SwissProt 中 200000 个非冗余的蛋白质组

表 10-2 InterPro（3.2 版）所综合的数据库资源

数据库	内容
Pfam7.3	3865 个结构域
PRINTS33.0	1650 个蛋白指纹
PROSITE17.5	1565 个谱图
ProDom2001.3	1346 个结构域
SMART3.1	509 个结构域
SwissProt40.27	113470 个记录
TIGRFAMs1.2	814 个蛋白家族
TrEMBL21.12	685610 个记录

注：以上资料来源于 http://www.ebi.ac.uk/interpro/release_notes.html（2002 年 10 月）。

织为 28000 个超家族，2600 种结构域，1300 种模体和 280 种翻译后修饰位点（Wu 等，2001）。iProClass 与其他 30 个数据库有链接。诸如 MetaFam、iProClass 和 InterrPro 这类资源对鉴定不同数据库之间的冲突，定义蛋白家族规模十分有用。

可以通过 http://www.ncbi.nlm.nih.gov/Entrez 访问 PopSet。

10.2.1.8 PopSet

下面我们转向一个多重比对的专业资源。PopSet（population data study sets）是 NCBI Entrez 站点内蛋白质、DNA 序列的比对集合，它们是来自种群、系统发生或突变研究的近缘序列。在 PopSet 中查询"hiv-pol"可以得到 300 个匹配。选择其中属于 Hermankova 等的记录，输出结果是一个来自多个病人的 pol DNA 序列比对（图 10-23）。

10.2.1.9 多序列比对编辑：手工与自动

有些数据库是手工编辑的，需要专家来进行注释。Sean Eddy 与其合作者们对 Pfam 进行了编辑，而 PROSITE 的编辑则是由 Amos Bairoch 等完成。BLOCKS 和 PRINTS 同样也是手工编辑的。显然专家注释的过程相当困难，但是这种方法在判断蛋白家族成员时有很大的优越性。像 DOMO、ProDom 和 MetaFam 这样的数据库，则采用自动注释；与第 5 章讨论的 PSI-BLAST 一样，比对错误与在原比对中加入一个新序列会带来很多问题。但是，自动注释对于分析大规模数据集（如从基因组计划中得到的上千条预测蛋白序列）是很有价值的。

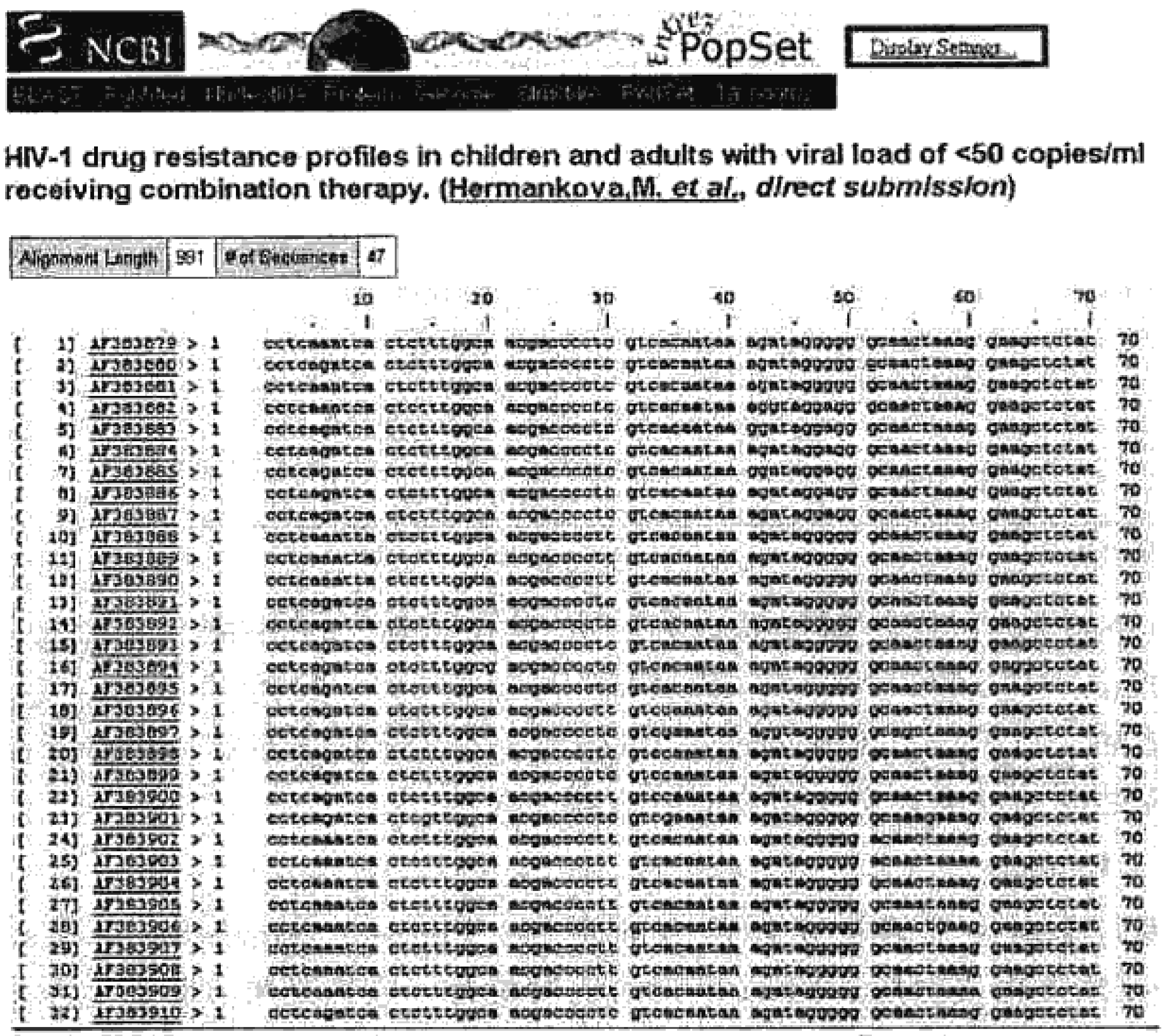
10.2.2 由用户产生的多序列比对

到目前为止，我们所讨论的都是包含了以多序列比对组织数据的蛋白家族数据库。尽管这些资源很有用，但有时候你得自己产生比对。利用表 10-3 中列举的站点和程序，你可以用一组序列产生比对。

10.2.2.1 ClustalW 和 ClustalX

ClustalW 是一个产生多重比对的网络工具（Thompson 等，1994），站点包括了一个用于提交序列的文本域，序列格式为 FASTA（图 10-1）。其他可以接受的格式有 NBRF/PIR、EMBL/SwissProt 和 GCG/MSF。以前我们曾看到过一个 ClustalW 的输出结果（图 10-2～图 10-5）。

(a)



(b)

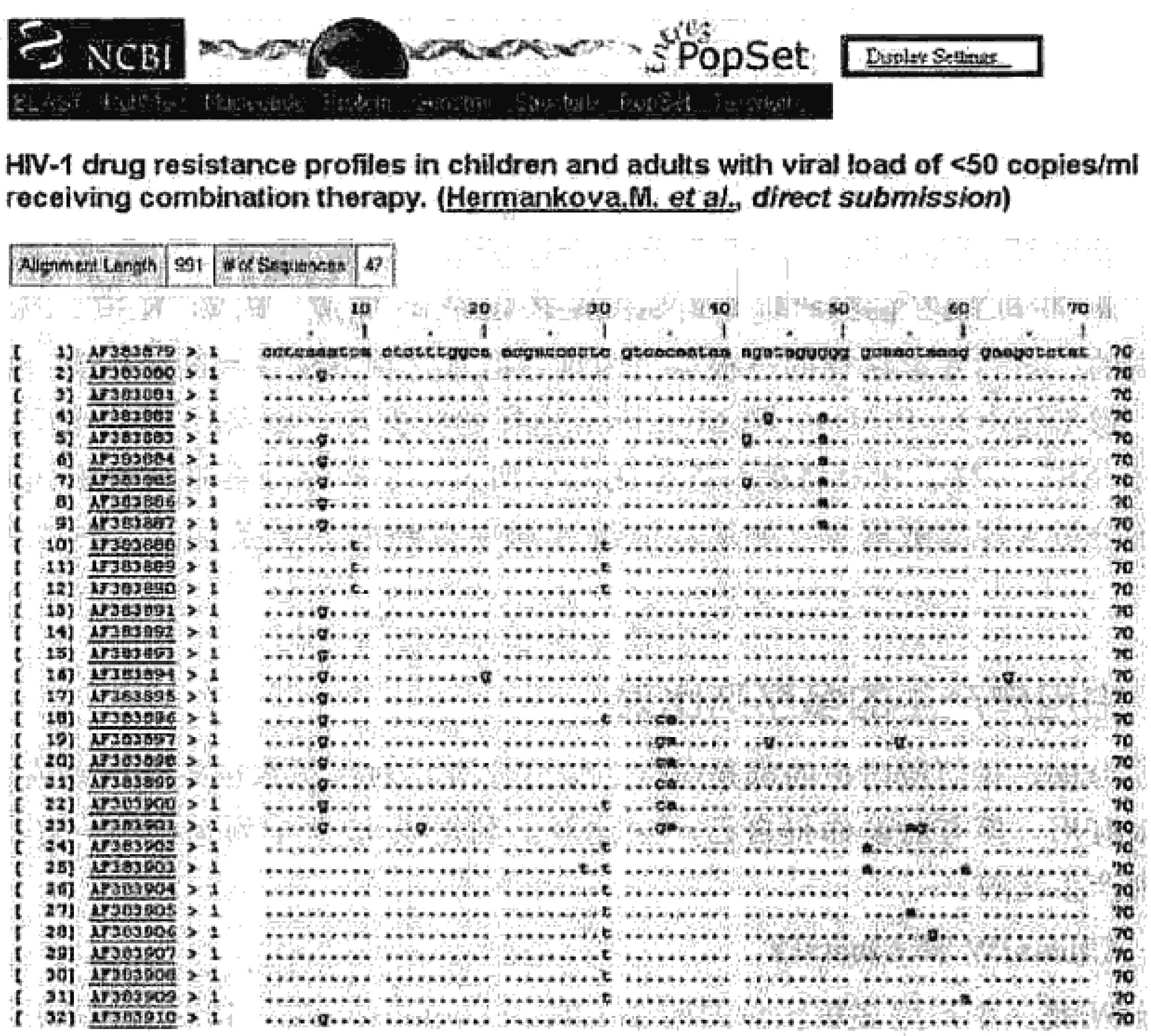


图 10-23 对 PopSet 的搜索得到的多重比对。点击“Display Setting”框，可以将结果保存为多种格式，如 FASTA 和 PHYLIP。这些对于进一步研究比对的系统发生关系是有用的。显示选项包括（a）全部序列，（b）变化位点

表 10-3 网上的多序列比对程序

程 序	描 述	网站地址
AMAS (Analyse Multiply Aligned Sequences)	位于 EBI，用于分析预先处理好的多重比对(MSA)	http://barton.ebi.ac.uk/servers/amas_server.html
CINEMA	用于多重比对的彩色互动编辑器	http://www.bioinf.man.ac.uk/dbbrowser/CINEMA2.1/
ClustalW	位于 EBI 和其他站点	http://www2.ebi.ac.uk/clustalw/
ClustalX	可用 FTP 下载	ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/
DIALIGN	对局部比对极为有用，位于德国 Bielefeld 大学	http://bibiserv.techfak.uni-bielefeld.de/dialign/
Match-Box Web Server 1.3	位于比利时 Namur 大学	http://www.fundp.ac.be/sciences/biologie/bms/match-box_submit.shtml
MultAlin	位于法国 Toulouse 的 INRA	http://prodes.toulouse.inra.fr/multalin/multalin.html
Multiple Sequence Alignment version 2.0	位于 GeneStream 服务器(http://www2.igh.cnrs.fr/home.eng.html)	http://xylian.igh.cnrs.fr/msa/msa.html
Multiple Sequence Alignment version 2.1	位于华盛顿大学	http://stateslab.wustl.edu/ibc/msa.html
Musca	IBM 的生物信息学小组	http://cbcsrv.watson.ibm.com/Tmsa.html
PileUp	从 SeqWeb 或 Genetics Computer Group(Unix 版本)获得	http://www.gcg.com
T-COFFEE	处理远缘蛋白时比 ClustalW 稍慢	http://www.ch.embnet.org/software/TCoffee.html

注：1. 更多算法参考 ExPAsy 的列表 (http://kr.expasy.org/tools/#align)。
2. MSA 表示多重比对。

ClustalX 是 ClustalW 的视窗界面（Thompson 等，1997）。这个程序功能强大、使用方便，而且已经被编译，可以下载到不同的平台上使用（见表 10-3 的 FTP 站点）。将一组载脂蛋白以 FASTA 格式保存为文本后，提交到 ClustalX [图 10-24(a)]，最初是不包含空白的。使用 "Do complete alignment" 命令可产生一个多重比对 [图 10-24(b)]。可以图形化形式对这个比对进行检查并修改，如根据先验信息调整空白的位置和其他参数。注意在图 10-24(b) 中有 2 个紧邻的空白区（箭头所指）。第三个序列中的 GQAFHL 可能是由于在此空白区发生插入引起的。类似的第四个序列中 RGVT 也可能是插入的结果。

Hall 曾经写过一部关于使用与系统发生软件联系的 ClustalW 的导论性教材（2001）。

10.2.2.2 PileUp (GCG)

GCG 包提供的 PileUp 是一个多重比对的程序。将 5 个载脂蛋白序列输入 GCG 的序列编辑器中：

① 来自大鼠（rat）的增味剂结合蛋白（odorant-binding protein，P08937，由鼻侧腺分泌进入黏液和眼泪的一种载脂蛋白）。

② 来自牛（bovine）的增味剂结合蛋白（odorant-binding protein，P07435，鼠 OBP 的直系同源物，具有相似的增味剂结合特性与三维结构。但是这两种蛋白质仅有 30%的序列相似性）。

③ aphrodisin，雌性大鼠（*Cricetus cricetus*）的阴道分泌物（I48075），是一种性外激素。

④ probasin，功能未知，前列腺上皮细胞的分泌蛋白（NP_061998）。

⑤ 牛（bovine）的头屑主要过敏原蛋白（dander major allergen，Q28133，Mantyjarvi 等，1996）。

将以上序列输入 GCG 的序列编辑器（Seqed）。可以通过建立文件的方法将这些序列组织起来。通过命令 "pileup@lipocalins" 使用 PileUp 程序将这组序列进行多重比对 [图

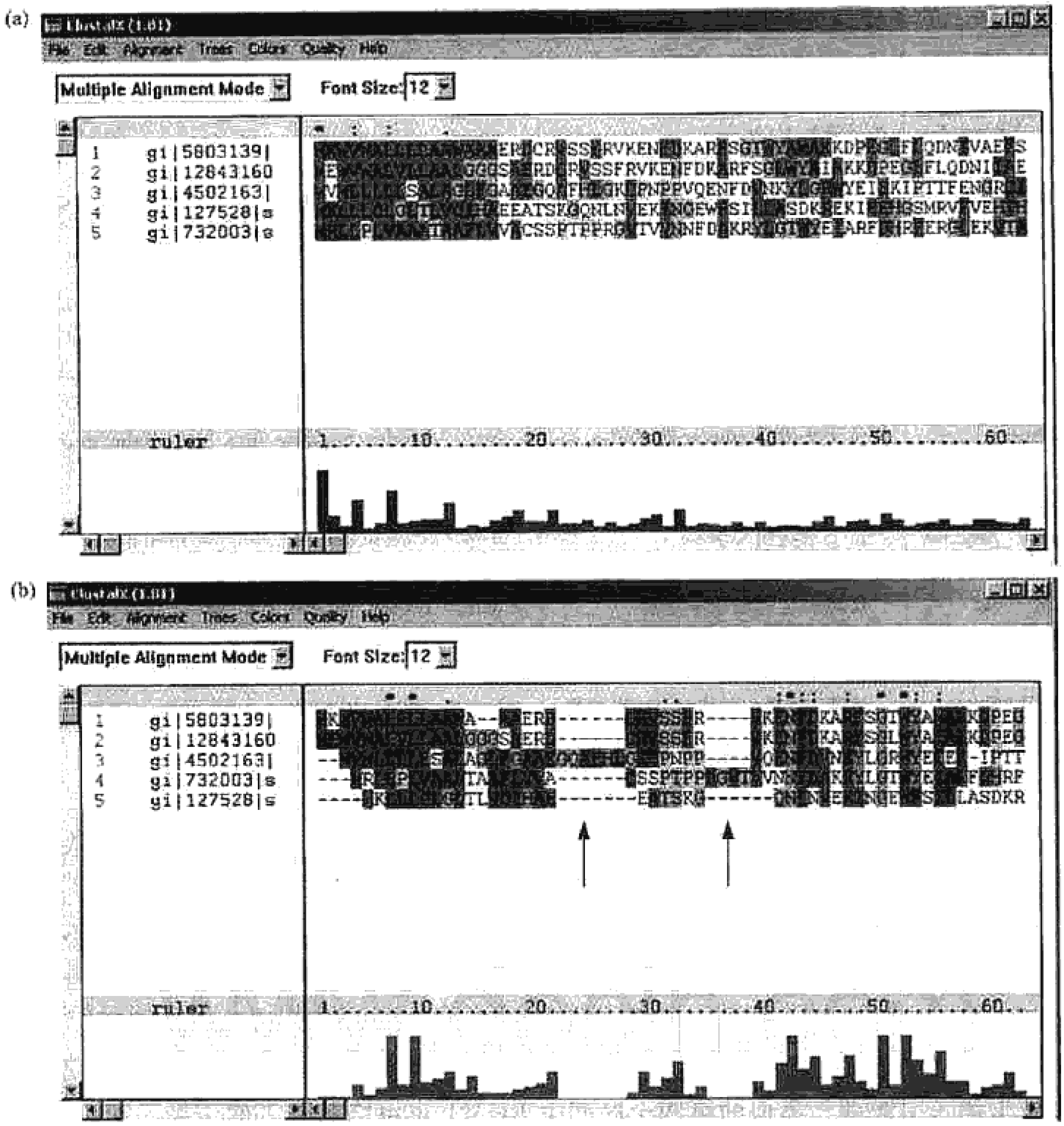


图 10-24 ClustalW 程序可以产生多重比对。从 NCBI Entrez 获得 5 个载脂蛋白序列的 FASTA 文件，保存为纯文本（见图 10-1），用 "load sequence" 命令将序列提交到 ClustalX。(a) 最初的序列是不包含空白的。(b) 使用 "Alignment" 下拉菜单中的 "Do Complete Alignment" 命令，产生一个多序列比对，可以进一步进行手工编辑。比对上方的星号表示 100% 的一致性，下方的直方图表示比对中残基保守性的相对量，ClustalX 比对结果可以用于系统发生分析（如 PAUP 等软件）

10-25(b)]。比对结果显示于图 10-25(a)。程序允许对空白罚分和空白延伸罚分进行调整，如果这些惩罚分值较小，在比对中会出现较多的空白 [图 10-26(b)]。作为一个生物学家，你需要对比对中是否插入或删除空白进行评估。在这个例子中，图 10-26(a) 的比对可能是最优的，因为保证了最关键的半胱氨酸和 GXW 区域的对齐。

> 如果可能，比较一下你所研究的蛋白质结构，这可以帮助你评估比对的质量。

还有很多选项可供使用（可以用 "genhelp pileup" 命令查看）。例如，比对中序列的排列顺序可以用命令 "nosort" 加以限制。默认的结果是，算法将把最优的匹配放在顶端（这里是老鼠 OBP 和大鼠的 aphrodisin）。PileUp 可以对最长为 7000 个残基或碱基的 500 个序列进行比对。

(a)

```
> cat>lipocalins

rnobp.pep
btobp.pep
ccaphr.pep
rnprob.pep
btdma.pep
```

(b)

```
> pileup @lipocalins

PileUp creates a multiple sequence alignment from a group of related
sequences using progressive, pairwise alignments.  It can also plot a
tree showing the clustering relationships used to create the alignment.

  1        rnobp.pep   172 aa
  2        btobp.pep   159 aa
  3       ccaphr.pep   167 aa
  4       rnprob.pep   177 aa
  5        btdma.pep   172 aa

 What is the gap creation penalty (* 8 *) ?

 What is the gap extension penalty (* 2 *) ?

 This program can display the clustering relationships graphically.
 Do you want to:

     A) Plot to a FIGURE file called 'pileup.figure'
     B) Plot graphics on HP7550 attached to /dev/tty15
     C) Suppress the plot

 Please choose one (* A *):

 The minimum density for a one-page plot is 4.7 sequences/100 platen units.
 What density do you want (* 4.7 *) ?

 What should I call the output file name (* lipocalins.msf *) ?

 Determining pairwise similarity scores...

  1   x   2      1.32
  1   x   3      1.89
  1   x   4      1.44
  1   x   5      1.48
  2   x   3      1.10
  2   x   4      1.16
  2   x   5      1.33
  3   x   4      1.48
  3   x   5      1.54
  4   x   5      1.48

 Aligning...

  1     ........-.
  2     ........-.
        ........-.
  3     ........-.
        ........-.
  4     .......-.

 FIGURE instructions are now being written into pileup.figure.

        Total sequences:          5
       Alignment length:        179
               CPU time:      00.27

  Output file:/nfs/groucho/user/pevsner/pevsner/bioinf.dir/lipocalins.msf
```

图 10-25 GCG 的 PileUp 程序可以进行多重比对，通过复制将序列粘贴到序列编辑器 Seqed 中。(a) 将需要比对的序列组织到一个文件（lipocalins）中。(b) 输入命令“pileup@lipocalins”运行程序。PileUp 允许改变空白罚分与空白延伸罚分，当降低罚分时，会在比对中引入更多的空白，远缘序列可能会更好地对齐。输出选项包括多重比对、导向树。和其他渐进比对程序一样，PileUp 也产生两两比对的相似性分值。输出结果是 FASTA 格式，其他程序［如 Pfam（本章）和 PAUP（第 11 章）］采用 WSF 作为输入输出格式

(a)

```
> cat lipocalins msf
!!AA_MULTIPLE_ALIGNMENT 1 0
PileUp of: @lipocalins

 Symbol comparison table: GenRunData:blosum62.cmp  CompCheck: 6430

                   GapWeight: 8
             GapLengthWeight: 2

 lipocalins.msf  MSF: 179  Type: P  September 2, 2001 17:46  Check: 6778

 Name: rnobp          Len:   179  Check: 2583  Weight:  1.00
 Name: ccaphr         Len:   179  Check: 6725  Weight:  1.00
 Name: btdma          Len:   179  Check: 4045  Weight:  1.00
 Name: rnprob         Len:   179  Check: 3848  Weight:  1.00
 Name: btobp          Len:   179  Check: 9577  Weight:  1.00


        1                                                  50
 rnobp  MVKFLLIVLA LGVSC.AHHE NLDISPSEVN GDWRTLYIVA DNVEKVAEGG
ccaphr  MVKILL..LA LVFSL.AHAQ DF....AELQ GKWYTIVIAA DNLEKIEEGG
 btdma  ~MKAVFLTLL FGLVCTAQET PAEIDPSKIP GEWRIIYAAA DNKDKIVEGG
rnprob  ~MRVILLLLT LDVLGVSSMM TDKNLKKKIE GNWRTVYLAA SSVEKINEGS
 btobp  ~~~~~~~~~~ ~~~~~~AQEE EAEQNLSELS GPWRTVYIGS TNPEKIQENG

        51                                                100
 rnobp  SLRAYFQHME CGDECQELKI IFNVKLDSEC QTHTVVGQKH EDGRYTTDYS
ccaphr  PLRFYFRHID CYKNCSEMEI TFYVITNNQC SKTTVIGYLK GNGTYQTQFE
 btdma  PLRNYYRRIE CINDCESLSI TFYLKDQGTC LLLTEVAKRQ EGYVYVLEFY
rnprob  PLRTYFRRIE CGKRCNRINL YFYIKKGAKC QQFKIVGRRS QD.VYYAKYE
 btobp  PFRTYFRELV FDDEKGTVDF YFSVKRDGKW KNVHVKATKQ DDGTYVADYE

        101                                               150
 rnobp  GRNVFHVLKK TDDIIFFHNV NVDESGRR.Q CDLVAGKRED LNKAQKQELR
ccaphr  GNNIFQPLYI TSDKIFFTNK NMDRAGQETN MIVVAGKGNA LTPEENEILV
 btdma  GTNTLEVIHV SENMLVTYVE NYD.GERITK MTEGLAKGTS FTPEELEKYQ
rnprob  GSTAFMLKTV NEKILLFDYF NRNRRNDVTR VAGVLAKGRQ LTKDEMTEYM
 btobp  GQNVFKIVSL SRTHLVAHNI NVDKHGQTTE LTELFVKLN. VEDEDLEKFW

        151                      179
 rnobp  KLAEEYNIPN ENTQHLVPTD TCNQ~~~~~~
ccaphr  QFAHEKKIPV ENILNILATD TCPE~~~~~~
 btdma  QLNSERGVPN ENIENLIKTD NCPP~~~~~~
rnprob  NFVEEHGIED ENVQRVMDTD TCPNKIRIR
 btobp  KLTEDKGIDK KNVVNFLENE DHPHPE~~~~
```

(b)

```
        1                                                  50
 rnobp  M VKFLLIVL  ALGV.SCAH HENLDISPSE VNGDWRTLYI VADNVEKVAE
ccaphr  M.VKILL..L  ALVF.SLAH AQ..DF..AE LQGKWYTIVI AADNLEKIEE
rnprob  MRVILLLLTL DVLGVSSMMT DKNL...KKK IEGNWRTVYL AASSVEKINE
 btdma  MKAVFL..TL .LFGLVCTAQ ETPAEIDPSK IPGEWRIIYA AADNKDKIVE
 btobp  ~~~~~~~~~~ ~~~~~~~~AQ EEEAEQNLSE LSGPWRTVYI GSTNPEKIQE

        51                                                100
 rnobp  GGSLRAYFQH MECGDECQEL KIIFNVKLDS ECQTHTVVGQ KHEDGRYTTD
ccaphr  GGPLRFYFRH IDCYKNCSEM EITFYVITNN QCSKTTVIGY LKGNGTYQTQ
rnprob  GSPLRTYFRR IECGKRCNRI NLYFYIKKGA KCQQFKIVGR RSQD.VYYAK
 btdma  GGPLRNYYRR IECINDCESL SITFYLKDQG TCLLLTEVAK RQEGYVYVLE
 btobp  NGPFRTYFRE LVFDDEKGTV DFYFSVKRDG KWKNVHVKAT KQDDGTYVAD

        101                                               150
 rnobp  YSGRNVFHVL KKTDDII..F .FHNVNVDES GRRQCDL..V AG...KREDL
ccaphr  FEGNNIFQPL YITSDKI..F .FTNKNMDRA G.QETNMIVV AG...KGNAL
rnprob  YEGSTAF.ML KTVNEKILLF DYFNRN....   RRNDVTRV AGVLAKGRQL
 btdma  FYGTNTLEVI HVSENMLVTY V...ENYD.. GERITKMTE. .G.LAKGTSF
 btobp  YEGQNVFKIV SLSRTHLVA.  .HNINVDKH G.QTTELTEL ...FVKLN.V

        151                                 188
 rnobp  NKAQKQELRK LAEEYNIPNE NTQHLVPTDT CNQ~~~~~~~
ccaphr  TPEENEILVQ FAHEKKIPVE NILNILATDT CPE~~~~~~~
rnprob  TKDEMTEYMN FVEEHGIEDE NVQRVMDTDT CPNKIRIR
 btdma  TPEELEKYQQ LNSERGVPNE NIENLIKTDN CPP~~~~~~~
 btobp  EDEDLEKFWK LTEDKGIDKK NVVNFLENED HPHPE~~~~~
```

图 10-26 用 PileUp 程序对 5 种载脂蛋白进行多重比对（见附录）。(a) 默认的结果。比对中有很多内部空白和末端空白。(b) 用较低的空白罚分（4）和空白延伸罚分（1）重复比对的结果。注意，比对中出现了更多空白。在进行系统进化分析时，要先删除模糊匹配区域

10.2.3 其他多序列比对软件

还有其他多序列比对软件，图 10-27 显示了一个概要（基于 Thompson 等，1999a）。这些程序包括 Dialign（Morgenstern，1999）、hmmt、Mlpima（Smith 和 Smith，1992）、Multal、Multalign、PRRP、SAGA（Notredame 和 Higgins，1996）、Sbpima（Smith 和 Smith，1992）。可以根据它们的适用范围（全局比对、局部比对）、在优化初始比对时采用的算法（渐进算法、迭代算法）进行分类。

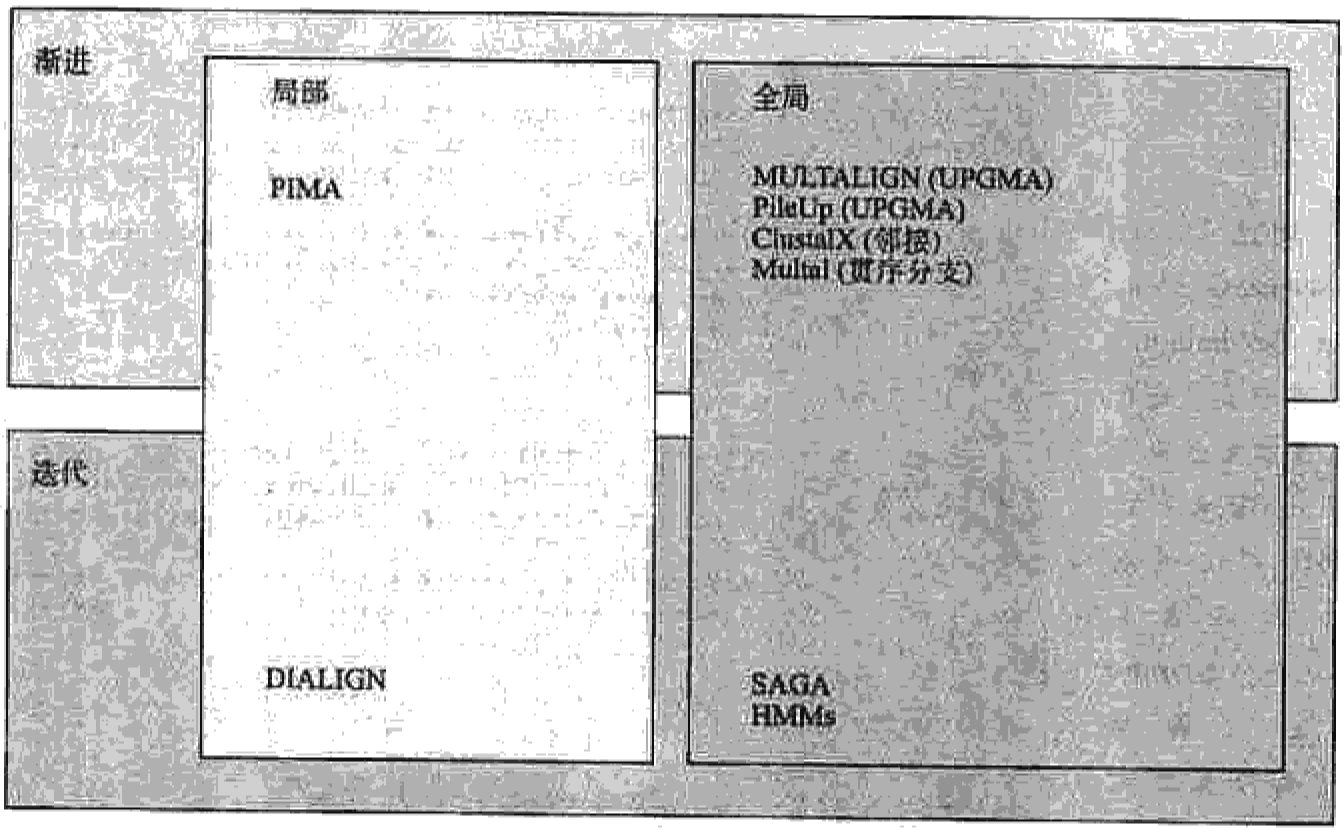


图 10-27 由 Thompson 等（1999a）总结的多重比对算法概要。渐进比对算法来自 Feng 和 Doolittle（1987）。迭代方法可以对初始比对进行优化，UPGMA 和邻接法与导向树的建立有关（第 11 章）

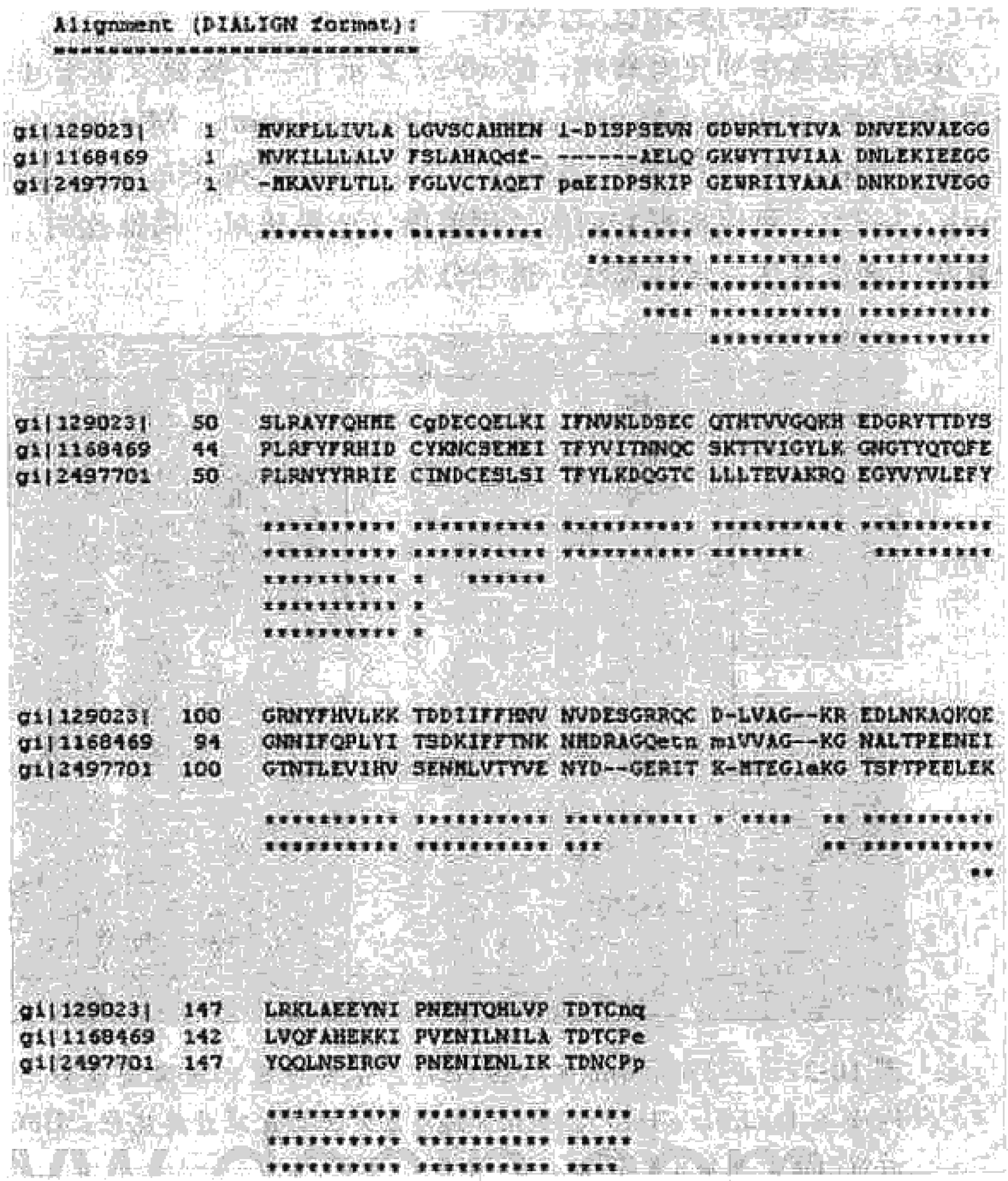
这些不同的程序并不是冗余的。用相同的序列，不同程序可能给出不同结果。此外，它们都具备很多特点使你能够对比对进行优化。例如 Dialign 和 Multalign 程序提供比对的健壮性度量（图 10-28 和图 10-29）。

10.2.4 多重比对算法的评估

几个小组已经系统地对多种多序列比对算法进行了精度和运行性质的评估（Morgenstern 等，1996；McClure 等，1994；Thompson 等，1999；Gotoh，1996；Briffeuil，1998）。具体方法是先得到（或创建）一个由已知结构蛋白质序列组成的参考比对数据库，接着用结构同源的序列测试比对程序。这样就能评估多重比对程序检测序列远缘关系的能力。Thompson 等（1999b）创建了 BAliBASE（142 个比对，超过 1000 个序列）作为参考比对数据库（http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE/index.html）。图 10-30 是 HIV-1 蛋白酶的比对。

通过这样的比较，可以得到下面的结论。

① 当比对序列的序列一致性降低时，比对的准确度下降（Briffeuil 等，1998）。对于一组相似性低于 25%的序列，问题就显得相当严重了。Thompson 等（1999a）发现对于相似性小于 25%的序列，最好的程序（PRRP，ClustalW 和 SAGA）能对齐 60%～70%的氨基



```
Alignment (DIALIGN format):
===========================

gi|129023|      1   MVKFLLIVLA LGVSCAHHEN 1-DISPSEVN GDWRTLYIVA DNVEKVAEGG
gi|1168469      1   MVKILLLALV FSLAHAQdf- -------AELQ GKWYTIVIAA DNLEKIEEGG
gi|2497701      1   -MKAVFLTLL FGLVCTAQET paEIDPSKIP GEWRIIYAAA DNKDKIVEGG

gi|129023|     50   SLRAYFQHME CgDECQELKI IFNVKLDSEC QTHTVVGQKH EDGRYTTDYS
gi|1168469     44   PLRFYFRHID CYKNCSEMEI TFYVITNNQC SKTTVIGYLK GNGTYQTQFE
gi|2497701     50   PLRNYYRRIE CINDCESLSI TFYLKDQGTC LLLTEVAKRQ EGYVYVLEFY

gi|129023|    100   GRNYFHVLKK TDDIIFFHNV NVDESGRRQC D-LVAG--KR EDLNKAQKQE
gi|1168469     94   GNNIFQPLYI TSDKIFFTNK NMDRAGQetn miVVAG--KG NALTPEENEI
gi|2497701    100   GTNTLEVIHV SENMLVTYVE NYD--GERIT K-MTEGlaKG TSFTPEELEK

gi|129023|    147   LRKLAEEYNI PNENTQHLVP TDTCnq
gi|1168469    142   LVQFAHEKKI PVENILNILA TDTCPe
gi|2497701    147   YQQLNSERGV PNENIENLIK TDNCPp
```

图 10-28 使用 Dialign（http://bibiserv.techfak.uni-bielefeld.de/dialign/）程序对3种载脂蛋白进行多重比对。这些蛋白质是大鼠（rat）的增味剂结合蛋白（P08937），仓鼠（hamster）的催情素（P09465），牛（bovine）的皮屑主过敏原（Q28133）。程序使用大写字母表示对齐的区域。5个星号表示完全保守的区域，鉴定出这些区域对于构建系统发生树很有用

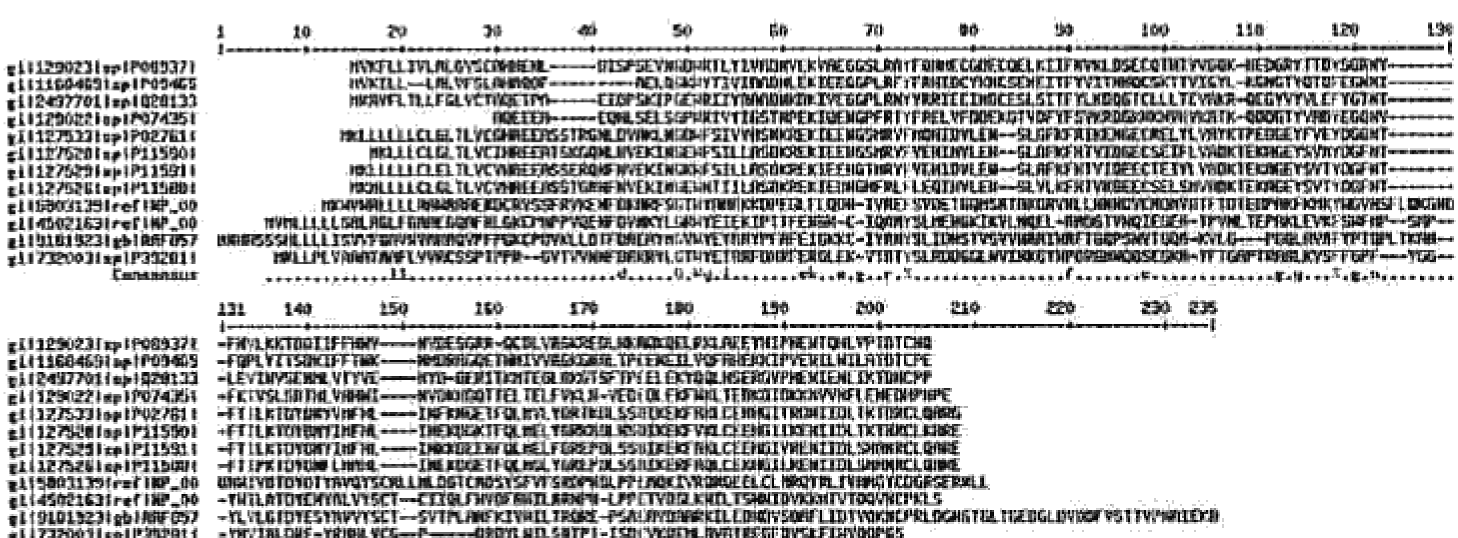


图 10-29 用 Florence Corpte 的 Multalin 程序进行多重比对（http://prodes.toulouse.inra.fr/multalin/multalin.html），对齐的列用红色（高保守）、蓝色（低保守）、黑色（中度保守）标记，程序提供了大量输入输出选项，如归类打分矩阵，空白罚分参数，配色方案等

```
Name                    hiv-1 protease

Number of sequences             4
Alignment Length                106
Longest Sequence                104
Shortest Sequence               98
Average Percent Identity        49
Maximum Percent Identity        86
Minimum Percent Identity        35

Sequence Name       SWISSPROT Accession
1fmb                    P32542
7upjB                   P03366
pol_sivcz               P17283
POL_SIVMK               P05897

Family    1fmb 7upjB pol_sivcz POL_SIVMK

1fmb          1   vTYNLEKRPTTIVLINDTPLNVLLDTGADTSVLTTahynrlkyrgrk.YQ
7upjB         1   pQFSLWKRPVVTAYIEGQPVEVLLDTGADDSIVAG.....iel.gnn.YS
pol_sivcz     1   pQITLWQRPLIPVKVEGQLCEALLDTGADDTVIER.....iqlqgl..WK
POL_SIVMK     1   pQFSLWRRPVVTAHIEGQPVEVLLDTGADDSIVTG.....iel.gph.YT

1fmb         50   GTGIGGVGGNVETFS.TPVTIKKKGRHIKTRMLVADIPVTILGRDILQDL
7upjB        44   PKIVGGIGGFINTLEYKNVEIEVLNKKVRATIMTGDTPINIFGRNILTAL
pol_sivcz    44   PKMIGGIGGFIKVKQFDNVHIEIEGRKVVGTVLVGPTPVNIIGRNILTQL
POL_SIVMK    44   PKIVGGIGGFINTKEYKNVEIEVLGKRIKRTIMTGDTPINIFGRNLLTAL

1fmb         99   GAKLV1
7upjB        94   GMSLN1
pol_sivcz    94   GCTLV.
POL_SIVMK    94   GMSLN1
```

Key

```
alpha helix          RED
beta strand          GREEN
core blocks          UNDERSCORE
```

你可以用 RSF 或具有特征表的 MSF 格式查看比对。

图 10-30 来自 BAliBASE（http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE/index.html）的一个多重比对。这个数据库包含了具有不同程度序列一致性的参考比对。图中是 HIV-1 蛋白酶的比对，一组相似性大于 35%的短序列。这个数据库中总共有 142 组比对，对评估 MSA 算法很有用

酸残基。当相似性较高（20%～40%）时，他们发现平均 80%的残基可以合理地对齐（Thompson 等，1999a）。

② 孤儿序列是指蛋白家族中高度分歧的成员。考察一个来自 10 个物种的 RBP 的多重比对，如果要在其中加入 OBP，OBP 会被视为孤儿序列。可能有人会认为孤儿序列会扰乱多重比对的组织，但实际上并不是这样。对于处理孤儿序列的情况，全局比对算法优于局部比对（Thompson 等，1999a）。

③ 几个多重比对可以联合起来，如一组近缘的 RBPs 和一组近缘的 OBPs。对这种情况，像 PRRP 和 SAGA 之类的迭代算法比 ClustalW 之类的渐进方法处理得好（Thompson，1999a）。但综合多种因素考虑，迭代方法相对渐进方法并没有明显优势。

④ 通常，一个家族内的某些蛋白质在氨基或羧基末端有较长的延伸。对这种情况局部比对程序的表现优于全局比对。一个例外是 PileUp，尽管采用全局比对，但是仍然表现得很好（Thompson 等，1999a）。总的来说，对于大多数多重比对，全局算法更为优越。

⑤ 在多重比对中采用不同的氨基酸替换矩阵的效果相对不明显（Gotoh，1996）。

10.3 展望

利用多重比对方法可以把一个蛋白家族（或 DNA 家族）的所有成员组织在一起。通过

多序列比对可以发现并定义任意基因或基因产物的旁系同源 paralogs（同一个物种内的家族成员）和直系同源 orthologs（种间的家族成员）。多重比对的蛋白家族数据库（如 Pfam 和 InterPro）规模增长相当快，作用也越来越重要。这些数据库通常都有很仔细的专家注释。一个大致的倾向是数据库将提供各种整合了的比对资源。随着基因组数据快速积累，已经必须使用像 InterPro 这样的资源对预测的蛋白序列进行功能注释。

有很多产生多重比对的程序。由于比对序列的一些特殊性质，这些程序经常给出不同的结果。通常合理的做法是使用 ClustalW、ClustalX 或 PileUp 中的一种来对齐你的序列，同时调整参数优化比对也是必需的。

10.4　常见问题

与多重比对相关的错误有很多，它们包括以下几个方面。

① 假阴性：比对时忽略了真正的同源。

② 假阳性：加入了不可靠的家族成员。

③ 比对错误：无法正确匹配远缘成员。

④ 比对错误：不恰当地加入空白。

10.5　网络资源

多序列比对被用来定义蛋白质的家系归属。在本章中，我们介绍了含有预先生成的比对的数据库（表 10-1）和允许用户进行自定义比对的程序（表 10-3）。

10.6　问题讨论

1. Feng 和 Doolittle 引入了“一旦是空白，永远是空白”的规则，也就是说在考虑空白的时候两个最近序列应该有最高的权重。为什么必须要引入这条规则？

2. 能否把 BLAST 搜索和 HMMs 联合起来？PSI-BLAST 和 Pfam 中基于 HMM 的搜索有什么不同？

10.7　问题

1. X 连锁的肾上腺白质退化症（X-ALD）是影响过氧化体（一种参与脂质代谢等活动的细胞器）功能的一种普通的遗传病。这种病是由 X 染色体 Xq28 位置编码 ALD 蛋白质（ALDP）的 *ABCD1* 基因突变引起。现在认为有 4 种 ALDP 相关蛋白位于人类的过氧化体：ALDP（NP _ 000024；745 个氨基酸残基）、ALDR（NP _ 005155，740 个残基）、PMP70（NP _ 002849，659 个残基）和 PMP70R（NP _ 005041，606 个残基）。两种酵母的 ADLP 相关蛋白也被鉴定：Pxa1p（NP _ 015178）和 Pxa2p（NP _ 012733）。这些蛋白质都属于一个更大的家族——ATP 结合盒转运蛋白，包括囊性纤维病跨膜转运调节物（CFTR）和多药物抗性蛋白（MDR）。用人类、老鼠和酵母的 ALDP 家族蛋白建立一个多重比对，找到可能的核苷酸结合位点 GPNGCGKS，这个区域的保守程度如何？

2. 非洲的白氏斑马现已灭绝，这是一种既像马又像斑马的动物。在 1872 年，人们拍到了最后一只白氏斑马的照片（图 10-31）。从博物馆的白氏斑马标本里人们得到了它的线粒体 DNA 并进行了测序。将白氏斑马（*Equus quagga boehmi*）、马（*Equus caballus*）和斑马（*Equus burchelli*）的 mtDNA 进行多重比对，白氏斑马与谁更接近？

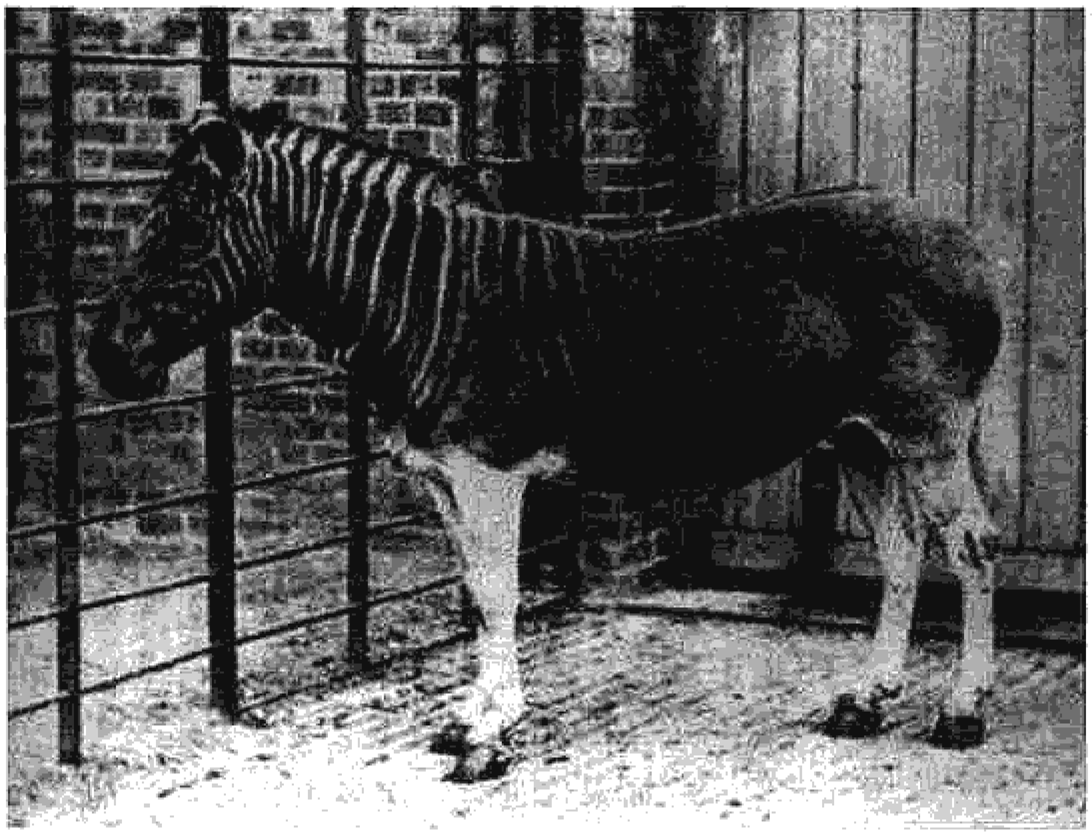


图 10-31 伦敦动物园白氏斑马的照片（1872）。这种已经灭绝的动物和马更接近，还是和斑马更接近？照片可以从以下地址获得：http://www.oursci.org/magazine/200108/010819.htm；homepages.tesco.net/～zoechaos/quaggas.htm；http://www.greenapple.com/～jorp/amzanim/quagga.htm；www.foek.hu/dodo/emlos/eqqu.htm

3. 为了确定哪种螺旋体侵人了严重牙周炎病人的牙龈，研究者们从龈下菌斑提取 DNA 和 RNA 样品并进行了克隆（Choi 等，1994）。在 NCBI Entrez 站点的 PopSet 中研究这些结果，确认你可以手工改善的一段比对区域。

4. 从关键字搜索开始，研究 Pfam 中的一个蛋白家族。

(a) 作为一个例子：查找 ABC 转运蛋白。

(b) 查看"top 20"的 Pfam 家族列表（http://Pfam.wustl.edu/browser.shtml）。

选择其中一个研究，尝试 JalView，用主成分分析或建树的方法查看家族成员。

5. 你确定了如下载脂蛋白的一些获取码：

人类视黄醇结合蛋白（human retinol binding protein，NP _ 006735）；

人类脂蛋白 D（human apolipoprotein D，NP _ 0016378）；

大鼠的增味剂结合蛋白（rat odorant-binding protein，P08973）。

将这些序列以 FASTA 格式贴到 ClustalW 中（http://www.ebi.ac.uk/clustalw/）建立多重比对。

10.8 自测题

1. 为什么 ClustalW（一个采用了 Feng 和 Doolittle 的渐进比对算法的程序）不报告 *E* 值？

(a) ClustalW 报告 E 值；

(b) ClustalW 使用的是全局比对，因此没有 *E* 值统计；

(c) ClustalW 使用的是局部比对，因此没有 *E* 值统计；

(d) E 值统计与多序列比对无关

2. 下列哪种程序不产生多序列比对？

(a) PSI-BLAST;

(b) ClustalW;

(c) PileUp;

(d) PHYLIP

3. Feng 和 Doolittle 方法的“一旦是空白，永远是空白”规则

(a) 保证空白不会因为在比对中加入序列而被不合理的填充；

(b) 假定在构建多重比对时应该给进化中早期分歧的序列较高的优先级；

(c) 假定最近序列的空白应该在多重比对中保留下来；

(d) 假定最远序列的空白应该在多重比对中保留下来

4. 下面哪一项陈述能最好地解释隐马模型（HMM）理论？

(a) 它依赖于一开始建立的系统发生树；

(b) 它计算各个位置每种氨基酸出现的概率；

(c) 它基于随机序列的分值计算多重比对；

(d) 它只能对齐属于已定义蛋白家族的序列

5. 下列哪种程序将查询蛋白和一组位点特异性记分矩阵进行比较？

(a) PHI-BLAST;

(b) RPS-BLAST;

(c) PROSITE;

(d) ProDom

6. 下列哪个数据库不是基于隐马模型（HMM）的？

(a) Pfam;

(b) PRINTS;

(c) SMART;

(d) TIGRFAMs

7. 搜索 InterPro 数据库的一个主要优点是

(a) SwissProt/TrEMBL 中几乎 25% 的序列在 InterPro 中有多于 1 个匹配；

(b) InterPro 成员数据库都用隐马模型来定义蛋白质家族；

(c) 在进行序列相似搜索时一个蛋白质可能有多于一个结构域，InterPro 数据库对于鉴定这些结构域很有用；

(d) 在进行序列相似搜索时，你可能要找查询序列中某个结构域的 reverse-profile HMM。InterPro 能够胜任这种工作

8. 如果你进行一次多重比对，当加入一个远缘蛋白（通常称作孤儿序列）时

(a) 它通常可以和其他蛋白对齐；

(b) 它通常不能和其他蛋白对齐

9. Pfam-A 和 Pfam-B 的主要区别在于

(a) Pfam-A 是手动编辑，而 Pfam-B 是自动编辑；

(b) Pfam-A 使用隐马模型，而 Pfam-B 不是；

(c) Pfam-A 提供全长的蛋白序列比对，Pfam-B 提供蛋白片段比对；

(d) Pfam-A 联合了来自 SMART 和 PROSITE 的数据，而 Pfam-B 没有

10.9 推荐读物

多重比对方面一项重要的工作是 Da-Fei Feng 和 Russell F. Doolittle（1987）发展了渐进比对方法。这项工作强调了蛋白质多重比对与其进化的关系。因此在第 11 章我们处理系统发生关系时采用了相关的方法。Doolittle（2000）也写了一些文章介绍了他在序列分析、系统发生、生物信息学方面的兴趣，以及他发展比对算法的历史。

在“Briefings in Bioinformatics”（2002 年 9 月第 3 卷 3 号）中有关于 InterPro 和它的组成数据库（PRINTS，PROSITE、Pfam，ProDom、SMART、TIGRFAMs、SwissProt 和

TrEMBL）的介绍。所有这些文章揭示了 InterPro 作为一个蛋白质家族和结构域综合资源的重要性。

一些研究团体已经系统地对各种多重比对方法进行了比较，阅读这方面的文献会使你对各种算法的优点和缺点有更深入的认识（Briffeuil 等，1998；Gotoh，1996；McClure 等，1994；Park 等，1998；Thompson 等，1999a）。Julie Thompson 等在 1999 年的工作尤其有价值。Park 等的文章（1998）对发掘 PSI-BLAST 和其他相关算法的运用十分有用。

在 "Methods in Enzymology"（1996 年第 266 期）中有一系列关于多重比对的文章（有 William Taylor、Da-Fei Feng、Russell Doolittle、Desmond Higgins 等）。

关于隐马模型 HMMs，本章中引用了不少导论文章，如 Ewan Birney 的一篇概要性综述（2001）。

参 考 文 献

Apweiler, R., et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29**, 37-40 (2001).

Attwood, T. K., et al. PRINTS prepares for the new millennium. *Nucleic Acids Res.* **27**, 220-225 (1999).

Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M. A. Hidden Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. USA* **91**, 1059-1963 (1994).

Bateman, A., et al. The Pfam protein families database. *Nucleic Acids Res.* **28**, 263-266 (2000).

Bateman, A., et al. The Pfam protein families database. *Nucleic Acids Res.* **30**, 276-280 (2002).

Birney, E. Hidden Markov models in biological sequence analysis. *IBM J. Res. Dev.* **45**, 449-459 (2001).

Briffeuil, P., et al. Comparative analysis of seven multiple protein sequence alignment servers: Clues to enhance reliability of predictions. *Bioinformatics* **14**, 357-366 (1998).

Choi, B. K., Paster, B. J., Dewhirst, F. E., and Gobel, U. B. Diversity of cultivable and uncultivable oral spirochetes from a patient with severe destructive periodontitis. *Infect. Immun.* **62**, 1889-1895 (1994).

Chothia, C., and Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826 (1986).

Doolittle, R. F. On the trail of protein sequences. *Bioinformatics* **16**, 24-33 (2000).

Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755-763 (1998).

Feng, D. F., and Doolittle, R. F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351-360 (1987).

Feng, D. F., and Doolittle, R. F. Progressive alignment and phylogenetic tree construction of protein sequences. *Methods Enzymol.* **183**, 375-387 (1990).

Gotoh, O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.* **264**, 823-838 (1996).

Hall, B. G. *Phylogenetic Trees Made Easy. A How-To for Molecular Biologists.* Sinauer Associates, Sunderland, MA, 2001.

Henikoff, S., Henikoff, J. G., and Pietrokovski, S. Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* **15**, 471-479 (1999).

Henikoff, S., Pietrokovski, S., and Henikoff, J. G. Superior performance in protein homology detection with the Blocks Database servers. *Nucleic Acids Res.* **26**, 309-312 (1998).

Hermankova, M., et al. HIV-1 drug resistance profiles in children and adults with viral load of <50 copies/ml receiving combination therapy. *JAMA* **286**, 196-207 (2001).

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27**, 215-219 (1999).

Krogh, A. An introduction to hidden Markov models for biological sequences. *In Computational Methods in Molecular Biology.* S. L. Salzberg, D. B. Searls, S. Kasif (eds.) Elsevier, New York, 1998, Chapter 4.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501-1531 (1994).

Mantyjarvi, R., et al. Complementary DNA cloning of the predominant allergen of bovine dander: A new member in the lipocalin family. *J. Allergy. Clin. Immunol.* **97**, 1297-1303 (1996).

McClure, M. A., Vasi, T. K., and Fitch, W. M. Comparative analysis of multiple protein-sequence alignment methods. *Mol. Biol. Evol.* **11**, 571-592 (1994).

Morgenstern, B. DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**, 211-218 (1999).

Morgenstern, B., Dress, A., and Werner, T. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci. USA* **93**, 12098-12103 (1996).

Mulder, N. J., et al. InterPro: An integrated documentation resource for protein families, domains, and functional sites. *Brief. Bioinform.* **3**, 225-235 (2002).

Mulder N. J., Apweiler R., Attwood T. K., Bairoch A., Barrell D., Bateman A., Binns D., Biswas M., Bradley P., Bork P., Bucher P., Copley R. R., Courcelle E., Das U., Durbin R., Falquet L., Fleischmann W., Griffiths-Jones S., Haft D., Harte N., Hulo N., Kahn D., Kanapin A., Krestyaninova M., Lopez R.,

Letunic I., Lonsdale D., Silventoinen V., Orchard S. E., Pagni M., Peyruc D., Ponting C. P., Selengut J. D., Servant F., Sigrist C. J., Vaughan R., Zdobnov E. M. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31,** 315–318 (2003).

Needleman, S. B., and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443–453 (1970).

Notredame, C. and Higgins, D. G. SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res.* **24,** 1515–1524 (1996).

Park, J., et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284,** 1201–1210 (1998).

Parry-Smith, D. J., Payne, A. W., Michie, A. D., and Attwood, T. K. CINEMA—a novel colour INteractive editor for multiple alignments. *Gene* **221,** GC57–63 (1998).

Ponting, C. P., Schultz, J., Milpetz, F., and Bork, P. SMART: Identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res.* **27,** 229–232 (1999).

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl. Acad. Sci. USA* **95,** 5857–5864 (1998).

Shoop, E., Silverstein, K. A., Johnson, J. E., and Retzel, E. F. MetaFam: A unified classification of protein families. II. Schema and Query Capabilities. *Bioinformatics* **17,** 262–271 (2001).

Sigrist, C. J., et al. PROSITE: A documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* **3,** 265–274 (2002).

Silverstein, K. A., Shoop, E., Johnson, J. E., and Retzel, E. F. MetaFam: A unified classification of protein families. I. Overview and statistics. *Bioinformatics* **17,** 249–261 (2001a).

Silverstein, K. A., et al. The MetaFam Server: A comprehensive protein family resource. *Nucleic Acids Res.* **29,** 49–51 (2001b).

Smith, R. F., and Smith, T. F. Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Eng.* **5,** 35–41 (1992).

Thompson, J. D., Higgins, D. G., and Gibson, T. J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).

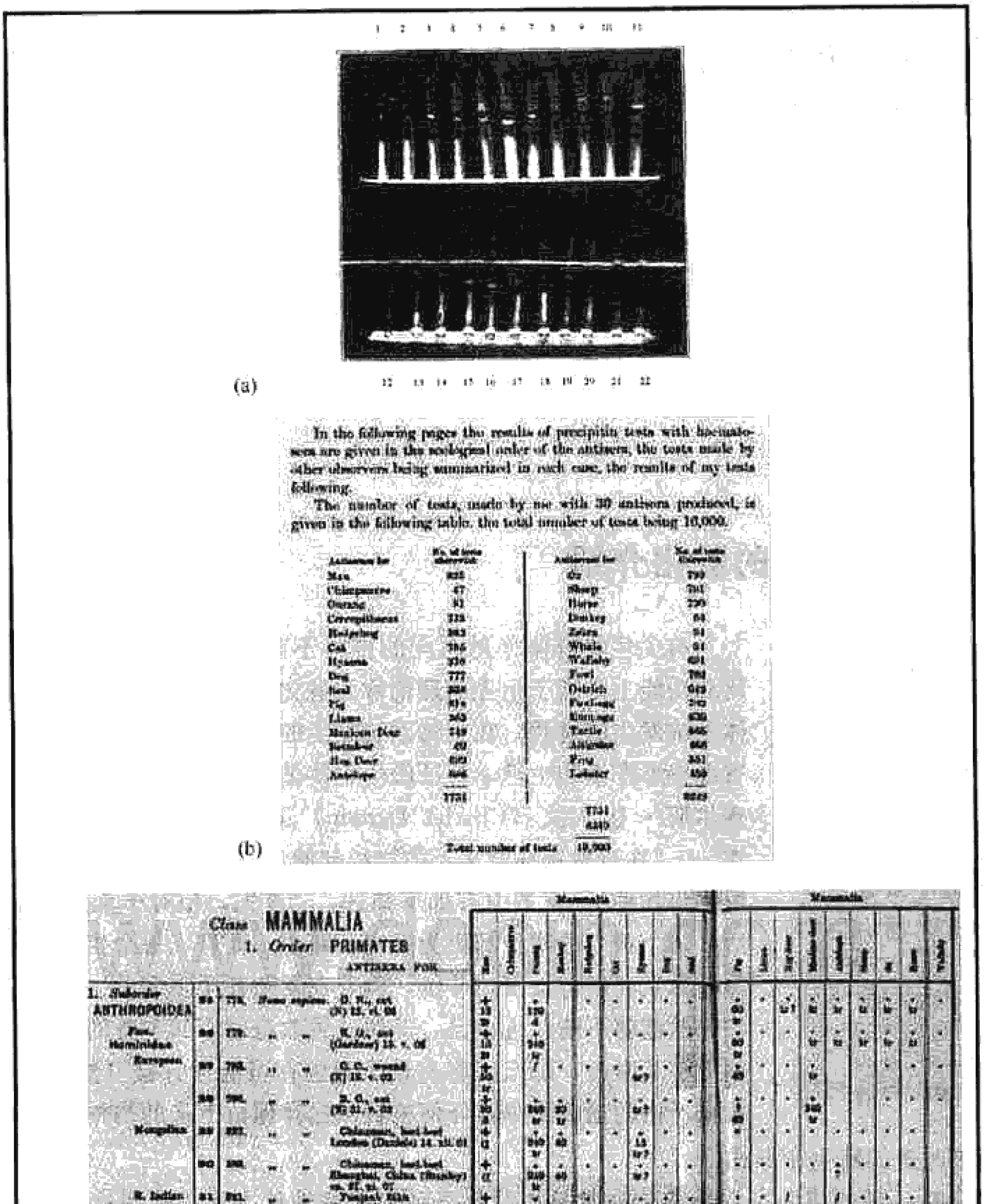
Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. The CLUSTAL X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25,** 4876–4882 (1997).

Thompson, J. D., Plewniak, F., and Poch, O. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* **27,** 2682–2690 (1999a).

Thompson, J. D., Plewniak, F., and Poch, O. BAliBASE: A benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* **15,** 87–88 (1999b).

Tuppy, H. Über die Artspezifität der Proteinstruktur. In A. Neuberger (ed.) Symposium on Protein Structure. John Wiley & Sons, New York, 1958, pp. 66–76.

Wu, C. H., Xiao, C., Hou, Z., Huang, H., and Barker, W. C. iProClass: An integrated, comprehensive and annotated protein classification database. *Nucleic Acids Res.* **29,** 52–54 (2001).



20世纪前半叶，基于分子数据的进化分析只有George Nuttall等发明的免疫沉淀测试。将抗毒素和来自不同物种的血清样品一起培养，记录沉淀所需时间和反应强度。(a) 反应试管 (Nuttall，1994)。(b) Nuttall实验记录（总共有16000个）的一部分（1904）。(c) Nuttall实验记录的第92页数据总结的一部分（1904）。900行的记录（这里显示了11行）包括了所有测试的血液样品，各列对应来自30个个体的抗毒素（这里显示18个）。数值代表反应所需时间（分钟）。符号表示反应程度（十表示强，·表示无）。字母D表示试管里出现沉淀。Nuttall利用这些数据来推断各种哺乳动物、鸟类、爬行类、两栖类和甲壳类的进化关系。20世纪五六十年代的进化分析中，氨基酸序列比对取代了免疫沉淀反应

第 11 章

分子水平的系统发生和进化

11.1 分子水平的进化介绍

我们将会在第 12 章中探索系统发生树的相关内容。
你可以在线阅读查尔斯·达尔文（Charles Darwin）的《物种起源论》，网址为 http://www.bbc.co.uk/education/darwin/origin/ 或 http://www.literature.org/authors/darwin-charles/the-origin-of-species/。

有关生物群体伴随时间的推移不断改变，从而导致其后代在结构和功能上不同于祖先的理论称为进化理论。进化也可被定义为生物体系统发生的过程。在生物系统发生的过程中，一种在遗传上得到具有独特形态学和生理学特征的生物体被定义为新物种。早在 1859 年，查尔斯·达尔文（Charles Darwin）就发表了其标志性的学术著作——《在自然选择意义下的物种起源》，亦可称为《在生存竞争中保留优势物种》。原书中有如下描述："首先，相对于无进化的永久存活，每个独立的物种都更为倾向于新生；而且，每个生物体每时每刻都在为生存进行反复的斗争，倘若在复杂甚至多变的生存条件下该物种仍然能够不断改进自己，即便这种改进对于其本身的有利影响可能是微乎其微的，那么其将有较大的生存可能性并被自然选择所保留。根据严格的遗传法则，任何被自然选择保留下来的物种都倾向于繁殖其已经被改进的新生命形式。"

进化是一种改进的过程。遗传通常是指后代子孙在某些性状方面仍然保持其父母的特征的现象，但在此过程中其身体的结构和功能已经发生了改变。

通常物种的改变符合如下三个基本机制（Simpson，1952）。

① 物种生长的影响条件得到发展。例如突发和疾病传染等引发的环境因子在自然界是不可继承的（尽管一个独立的个体对疾病或环境的刺激通常能够被控制在某个特定程度上，相关内容将在第 18 章中予以讨论）。

② 有性繁殖的机制必定在世代交替中有所改变。基因将在一个特定的染色体联会过程中发生重排，这样才能确保其子孙同时得到分别来自双亲的染色体遗传物质。

③ 突变能够对基因产生改变并且在染色体中较为频繁发生。

在分子水平上，进化是一种伴随着突变的自然选择过程。分子进化理论着重于研究不同系统发生树分支上基因和蛋白质的变化方式。这个学科也常常使用现今的生物体数据去重建

进化历史上曾经出现过的物种。系统发生关系是表示物种进化关系的参考依据。从传统意义上来说，物种的系统发生关系是通过比较大量不同的生物体形态学特征的差异来估算的（Mayr，1982）。但是，分子水平的序列数据也能被用于物种系统发生关系分析的研究中。通常研究者们用树的形式来表示不同物种间的进化关系，这为之前可能发生的生物学事件提供了假说。

"phylogeny"（系统发生）这个单词来源于希腊语"phylon"（表示种族、谱系、类别的意思）和"genesia"（表示起源的意思）。第12章卷首插图上的 Ernst Haeckel 进化树包含了系统演化、分类和生态学的多种信息。他同时提出"个体的发生是一种简便和快速的系统发生的重演，它是由生物体遗传（generation）和适应环境（maintenance）共同作用得到的生理学功能所决定的"（Haeckel，1900）。详情可参见 http://www.ucmp.berkeley.edu/history/haeckel.html。

11.1.1 历史背景

从20世纪50年代开始研究胰岛素至今，科学家们对分子进化的研究已经有了巨大的飞跃。胰岛素是一种刺激生物体吸收葡萄糖的小分子蛋白，一般情况下，待吸收的葡萄糖与胰腺腺泡细胞上的胰岛素受体相结合。早在1953年，桑格（Frederick Sanger）研究小组就采用实验手段确定了胰岛素的氨基酸序列，这是人类第一次对某种特定蛋白质的体外完全测序。成熟的、具有生物学活性的胰岛素蛋白质由2个亚基组成，这两个亚基分别称作A链和B链，它们是通过分子间的二硫盐桥键共价结合在一起的。最近的研究表明，人类的胰岛素蛋白分子是由一条信号肽、一条B链、一段被称作C链的多肽和一条A链共同组成的［图11-1(a)］。其中C多肽的两端各有一对二元残基［分别为精氨酸-精氨酸或者赖氨酸-精氨酸；详见图11-1(a)和图11-1(b)］作为它的蛋白酶切位点。

Frederick Sanger"因其在蛋白质结构研究上的突出贡献，特别是在胰岛素研究方面"而获得了1958年的诺贝尔化学奖（详见 http://www.nobel.se/chemistry/laureates/1958/）。1980年他又"因为其在核苷酸测序工作方面的突出贡献"而与 Paul Berg 和 Walter Gilbert 一起分享了该年的诺贝尔化学奖。

桑格（Sanger）和其他研究者对五个物种（母牛、绵羊、猪、马和鲸鱼）的胰岛素蛋白进行了测序。结果非常清晰和直接地反映了A链和B链在不同物种中表现出高度的保守性。此外，氨基酸的差异性只严格存在于3个碱基范围内，并且位于A链的一个二硫盐桥环区域中［详见图11-1(b)阴影部分］。这种结果暗示氨基酸替代事件的发生是非随机性的，某些改变是能明显影响生物学活性的，另一些改变是根本没有意义的（Anfinsen，1959）。在二硫键环内的差异性就是不具有生物学意义的一种，通常被称为中性的改变（Jukes和Cantor，1969；Kimura，1968）。然后，当我们比较具有较少生物学重要性的C多肽和具有较多生物学活性的A链和B链序列时，发现两者在不同物种间的序列保守性存在着惊人的差异性，其中C多肽在不同物种间的序列保守性要低得多。Kimura（1983）研究发现C多肽的进化速率为 2.4×10^{-9} 个氨基酸位点/年，大约为A链和B链进化速率（0.4×10^{-9} 个氨基酸位点/年）的6倍。在核苷酸水平上，编码生成C多肽的DNA的进化速率也比编码A链和B链的DNA快将近6倍左右（Li，1997）。

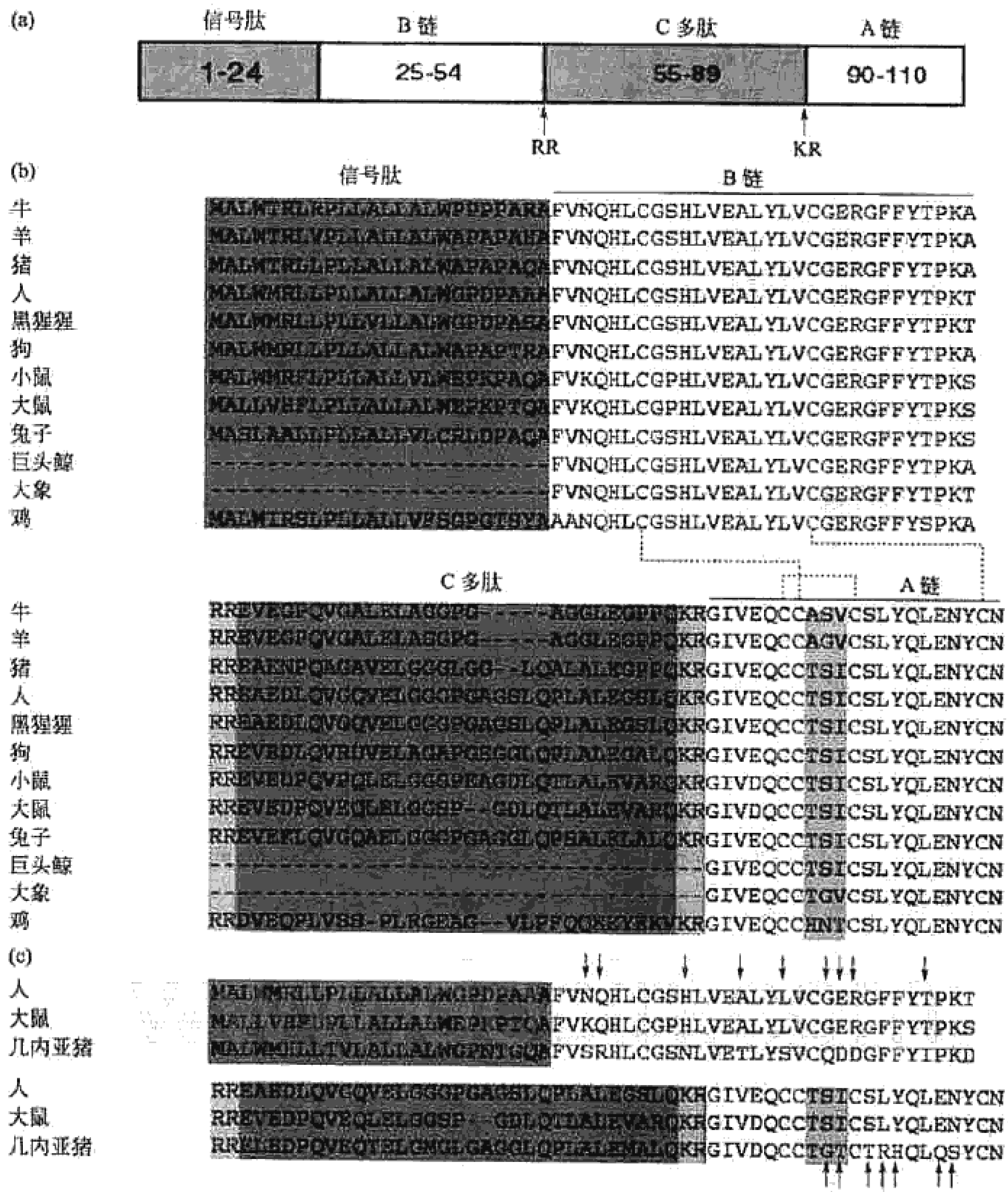


图 11-1 从1956年开始的有关胰岛素的研究能更好地帮助我们理解分子进化。(a) 人类的胰岛素分子由一个信号肽片段（用于细胞内的转运：一般为1～24个氨基酸残基）、一段B链、一段C多肽和一段A链组成。C多肽的两端分别有一个二元碱基对（RR、KR）组成的蛋白酶剪切位点。成熟胰岛素蛋白的A链和B链之间是通过共价的二硫盐桥键来连接的。(b) 12个物种的胰岛素蛋白多重序列比对图。替换事件的发生是遵循非随机模式的。值得注意的是胰岛素A链部分氨基酸残基在不同的物种间几乎是完全保守的，除了图中阴影标注出来的3个氨基酸（可见这3个氨基酸都是位于A链的二硫键形成的环内部）。但是，C多肽的氨基酸替换事件的发生频率要比A链和B链高将近6倍（Kimura，1983），图中多重序列比对显示出来的空白间隙（gaps）就是证据之一。半胱氨酸之间的二硫盐桥键是通过虚线来表示的，其中各个物种的登记序列分别是：NP _ 000198（人）、P30410（黑猩猩）、NP _ 062002（大鼠）、P01321（狗）、NP _ 032412（小鼠）、P01311（兔子）、P01315（猪）、P01332（小鸡）、NP _ 776351（牛）、P01318（羊）、INEL（大象）和INWHP（巨头鲸）。(c) 几内亚猪的胰岛素蛋白序列（Cavia porcellus，登记号为P01329）的进化速率大约比其他物种快7倍左右。对人、小鼠和几内亚猪的胰岛素序列进行比对。图中箭头标示出在16个氨基酸位置上几内亚猪的序列与人和/或者小鼠存在差异

由于其他物种的胰岛素序列的测定，一个惊人的发现显露出来。几内亚猪和一种与 *Caviidae*（coypu：南美洲产的一种水栖动物）存在高度近缘关系的物种体内的胰岛素的进化速率比其他物种快将近 7 倍左右。从图 11-1 可见，几内亚猪的胰岛素蛋白氨基酸序列的 A 链和 B 链部分有 16 处与人和小鼠不同的氨基酸位点。关于这个现象的解释在 1979 年由 Jukes 提出：因为几内亚猪和 coypu 的胰岛素蛋白没有与 2 个锌原子相结合，而其他物种的胰岛素蛋白都有。这就暗示含锌的络合物可能对大多数的胰岛素蛋白的氨基酸残基具有较强的功能性约束。几内亚猪和 coypu 的胰岛素蛋白序列所受约束较少。

早在 20 世纪 60 年代，另外一个实验室的研究者对催产素（oxytocin）和后叶加压素（vasopressin）进行测序时发现：虽然这两个多肽片段只在两个氨基酸位置上存在差异性，却有着截然不同的生物学功能（图 11-2）。1960 年 Max Perutz 和 John Kendrew 分别解决了血红蛋白（hemoglobin）和肌红蛋白（myoglobin）的结构问题。这两个蛋白都是氧分子载体，并且分享相关的结构和具有一定的同源性。因此 20 世纪 60 年代人们开始较为清晰地认识到蛋白质的氨基酸序列对其具有生物学意义的结构和功能有重大影响。

Perutz 和 Kendrew "因为其在球状蛋白结构研究上的突出贡献"而获得 1962 年的诺贝尔化学奖。详情可参见：http://www.nobel.se/chemistry/laureates/1962/。

催产素	CYIQNCPLG
后叶加压素	CYFQNCPRG

图 11-2　催产素（oxytocin：LocusID 5020；NP _ 000906）和后叶加压素（vasopressin：LocusID 551；NP _ 000481）只在两个氨基酸位置上存在差异性，但是它们却有着截然不同的生物学功能。这个早在 20 世纪 60 年代的序列比对的结果导致人们对氨基酸序列决定蛋白质功能的认可度产生了怀疑

11.1.2　分子时钟假说

早在 20 世纪 60 年代，原始的氨基酸序列数据得到较为充分的积累，例如不同物种间的可溶性蛋白（血红蛋白、细胞色素 C 和血纤维蛋白酶等）的原始数据。某些蛋白进化得非常慢，例如不同生物体内得到细胞色素 C；而其他蛋白家族却在短期内积累了大量的氨基酸替换事件。Linus Pauling 和其他研究人员因此提出了一种分子时钟的假说（Zuckerkandl 和 Pauling，1962；Margoliash，1963）。这种假说认为对于每一个给定的基因（或者蛋白质），其分子进化速率大致是恒定的。

1971 年 Richard Dickerson 对分子时钟的存在性进行了研究证明（图 11-3）。他一共分析了 3 种能够得到足够序列数据的蛋白质：细胞色素 C、血红蛋白和纤维素蛋白多肽。对于每个蛋白质，他都将两个物种间具有差异的氨基酸位点数目和物种离散所需要的时间（以百万年为单位来计时，MY）关系进行作图。这些物种的离散所需时间都是从古生物学中得到的。这些坐标图的 *y* 坐标是由修正后的每 100 个氨基酸残基中所发生的改变的数目来量度的，记作 *m*。*m* 值计算方法如下：

$$\frac{m}{100}=-\ln\left(1-\frac{n}{100}\right) \tag{11-1}$$

这个方程也可以改写成如下形式：

$$\frac{n}{100}=1-e^{-(m/100)} \tag{11-2}$$

式中，m 是一个蛋白质中每 100 个氨基酸片段中发生改变的所有氨基酸数目；n 是每 100 个观察得到的氨基酸残基数目。这个修正可以调整直接观察所不能发现的却实际存在的氨基酸残基变化，例如在同一个位点上的两个或多个氨基酸的改变（图 11-11）。

从这些直线（图 11-3）的特征上可得出如下的一些结论（Dickerson，1971）。

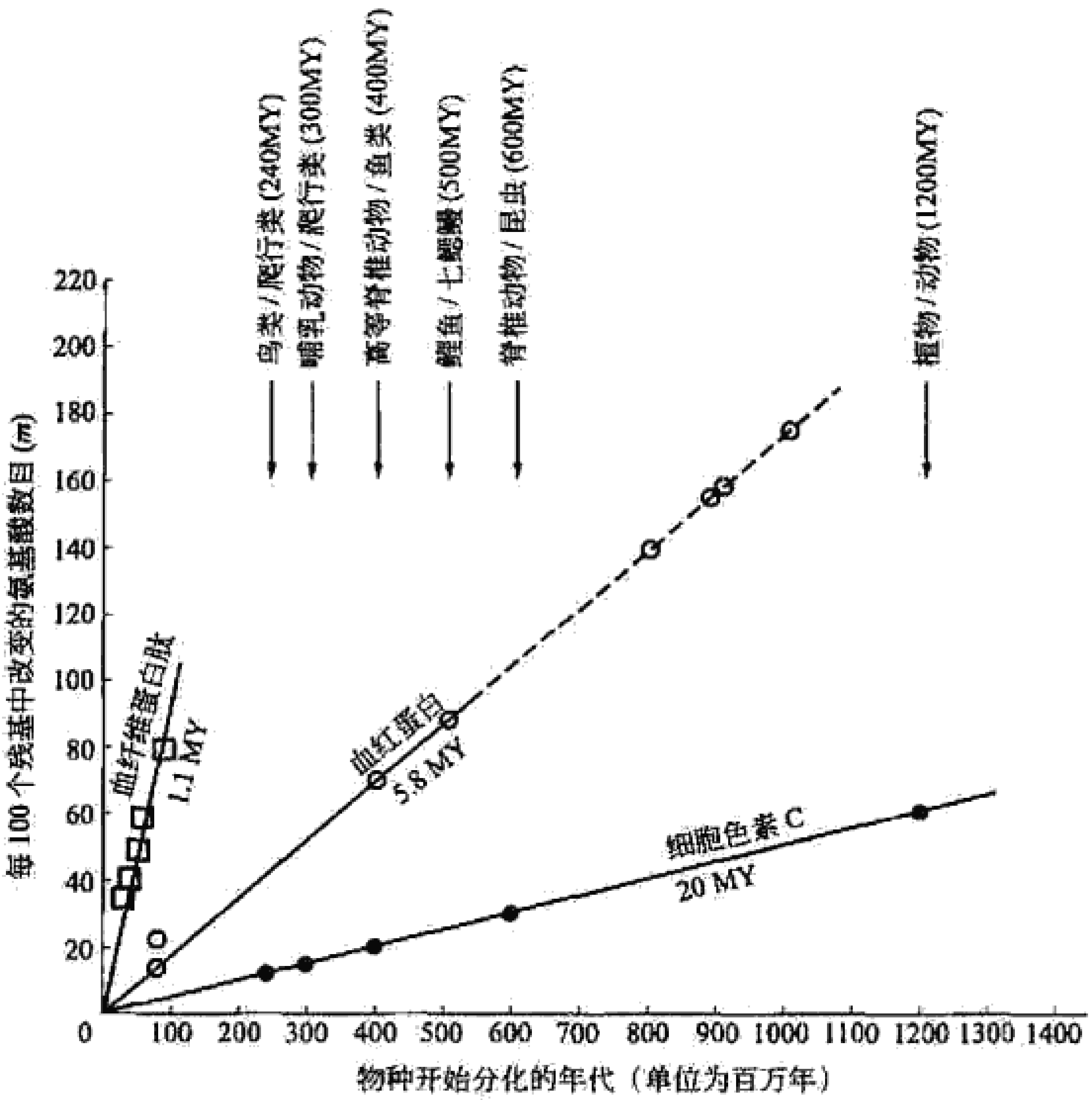


图 11-3 不同蛋白质间氨基酸数目改变的比较（y 轴）和物种开始在进化树上发生分支的时间（x 轴）共同显示了单个蛋白都是以各自独特的速率来进化的。某些蛋白（例如来源于不同生物体的细胞色素 C）进化较为缓慢；其他蛋白（例如血红蛋白）以一个中间的速率进化；血纤维蛋白肽的替换速度最快。这种现象被 Pauling Margoliash 和其他人在 20 世纪 60 年代用分子时钟的假说来进行描述。不同生物的分化时间（用箭头标示）主要用化石证据来估算。MY 表示过去的百万年（Dickerson 于 1971 年提出）；一些数据点和标准差的度量方式按照使用惯例被省略。数据的使用经过授权

① 对于每一个蛋白质，所得的数据都位于一条直线上，这就暗示了每个蛋白质的替换速率是恒定的。

② 每个蛋白质的平均替换速度是截然不同的，例如血纤维蛋白肽有相对较高的替换速率。在进化树的分叉中每 1% 的氨基酸序列发生改变所需要的时间（用百万年的量纲 MY 来度量）分别是：细胞色素 C 为 20.0MY，血红蛋白为 5.8MY，血纤维蛋白肽为 1.1MY。

③ 蛋白质家族中所遵循的变化率反映了自然选择对其功能性改变的约束力。

替换速率是通过测定每个位点每年所发生的替换的数目来衡量的，记作 λ。其中部分 λ 值在表 11-1 中给出。值得注意的是组蛋白和泛素的替换率非常低。作为参考，表 11-2 列出了由交互式 BLAST 搜索得到的人类、线虫（*Caenorhabditis elegans*）和酵母（*Saccharomyces cerevisiae*）蛋白质组中迄今为止所知的 20 种最为保守的蛋白质。

值得注意的是，我们虽然说组蛋白的替换率（substitution）很低，但是并没有说其突变率（mutation）很低。突变是一种生物化学过程，其导致的结果是最终在序列上发生改变。例如，一种 DNA 聚合酶以一种特殊的突变率（mutation）来复制 DNA。而替换率（substitution）是我们所观察到的在氨基酸和蛋白质序列中的改变（例如发生在不同组蛋白之间的）。观察到的替换速率同时反映了突变和选择，在这个进化过程中某个特征被选择（或者被放弃）。

表 11-1 不同物种中每 10^9 年每个氨基酸位点发生的替换率（$10^9\lambda$）的值

蛋白质	进化速率	蛋白质	进化速率
血纤维蛋白肽	9.0	促甲状腺素 β 链	0.74
生长激素	3.7	甲状旁腺(激)素	0.73
免疫球蛋白 κ 链 C 区域	3.7	小清蛋白	0.70
κ 酪蛋白	3.3	胰蛋白酶	0.59
免疫球蛋白 γ 链 C 区域	3.1	促黑细胞激素 β	0.56
前垂体激素 β 链	3.0	α 晶状体球蛋白 A 链	0.50
免疫球蛋白 λ 链 C 区域	2.7	安多酚	0.48
乳白蛋白	2.7	细胞色素 b_5	0.45
表皮生长因子	2.6	胰岛素(除了豚鼠和南美洲一种水栖动物的)	0.44
somatotropin	2.5	降血钙素	0.43
胰腺核糖核酸酶	2.1	(垂体)后叶激素运载蛋白 2	0.36
血清白蛋白	1.9	质体蓝素	0.35
磷脂酶 A2	1.9	乳酸脱氢酶	0.34
泌乳刺激素	1.7	腺苷酸激酶	0.32
碳酸酐酶 C	1.6	细胞色素 C	0.22
血色素 α 链	1.2	肌钙蛋白 C，骨骼肌	0.15
血色素 β 链	1.2	α 晶状体球蛋白 B 链	0.15
胃泌激素	0.98	胰高血糖素	0.12
溶解酵素	0.98	谷氨酸脱氢酶	0.09
肌血球素	0.89	组蛋白 H2B	0.09
淀粉体 AA	0.87	组蛋白 H2A	0.05
神经生长因子	0.85	组蛋白 H3	0.01
酸蛋白酶	0.84	泛素(ubiquitin)	0.01
髓磷脂基础蛋白	0.74	组蛋白 H4	0.01

注：1. Dayhoff 在 1978 年提出了这个在大约 1 亿年的进化历程中每 100 个氨基酸残基中发生的可接受的点突变率概念（PAMs）（进一步资料可与框 3-3 做比较来得出），其中可接受的血浆血清蛋白在每亿年进化历程中每 100 个氨基酸残基发生的点突变率的值为 19 PAMs。

2. 数据来源于 Dayhoff（1978），Nei（1987）。数据的使用经过授权。

表 11-2 线虫（*Caenorhabditis elegans*）、酵母（*Saccharomyces cerevisiae*），以及人类中 20 种最为保守的蛋白质列表

蛋白质	成对确定的百分比		
	线虫/人类	线虫/酵母	酵母/人类
1. H4 组蛋白	99	91	92
2. H3.3 组蛋白	99	89	90
3. 激动素 B	98	88	89
4. 泛素	98	95	96
5. 钙调蛋白	96	59	58
6. 胰蛋白酶 2	94	75	76
7. 泛素结合酶 UBC4	93	80	80
8. 网格衣被联合蛋白	93	48	48

续表

蛋白质	成对确定的百分比		
	线虫/人类	线虫/酵母	酵母/人类
9. 胰蛋白酶	93	73	74
10. 二磷酸腺苷酸(ADP)糖基化因子1	93	77	77
11. 动力蛋白轻链1	92	51	50
12. 三磷酸鸟苷(GTP)结合核蛋白 RAN	89	82	81
13. 色氨酸/苏氨酸蛋白质磷酸(酯)酶 PP1γ	89	84	85
14. 色氨酸/苏氨酸蛋白质磷酸(酯)酶 PP2β	89	74	76
15. 泛素结合酶 UBE2N	88	67	70
16. 组蛋白 H2A.Z	88	69	69
17. 组蛋白 H2A.2	87	79	76
18. DIM1p 同源蛋白	86	61	65
19. G25K 三磷酸鸟苷(GTP)结合蛋白	86	76	80
20. 40S 核糖体亚基蛋白 S15A	86	76	77

注：数据来源于1999年 Peer Bork 和他的一些同事（Copley 等）的研究工作，他们将当时可获得的所有这三个不同物种的完整蛋白质组的数据进行了交互式的 BLAST 搜索。相关内容将在第16章中有详细记载，当然，需要额外说明的是，数据注释的进展也是日新月异的。数据的使用经过授权。

分子时钟假说中一个隐含的重大意义就是：如果蛋白质序列的进化保持一个恒定的速率，那么这个速率就能被用于推算不同物种序列发生分化的时间。通过这种方式一些不同物种的系统发生关系就能被确定下来。这就类似于用测定放射元素半衰期的方法来鉴定地质标本的所属年代。一个有关分子时钟将如何被使用的例子将在框11-1给出。

分子时钟假说并不能应用于所有的蛋白质，其中有大部分的反例和警告被人们所关注到：

① 不同的生物体中的分子进化速率差别是很大的。例如，某些病毒相对于其他生命形式来说，其序列的变化速率要快得多。

② 不同基因的分子钟一般不同，甚至在单个基因的不同部分有时也不同（例如图11-1中的胰岛素就是一个最好的例子）。分子时钟的主要推动压力是自然选择。啮齿动物的分子时钟比灵长类的动物要快，这是因为他们世代间的间隔时间较灵长类动物短，且具有很高的代谢速率。

③ 分子时钟只能用于在长期进化历程中仍保持其生物学功能的那些基因。基因可能会因为进化而变得没有生物学功能（例如成为假基因）或者在自我扩增后改变了本身的生物学功能。如这样，其氨基酸的序列改变速度就会相对较快。

我们将会在第15章对整个基因组的自我扩增所伴随的后续的、快速的突变和基因丢失现象展开深入讨论（研究对象为酵母：*Saccharomyces cerevisiae*）。

④ 进化的速率是可变的，某些基因在发生自我扩增以后进化速度会大大提高。例如，基因自我扩增以后产生了α血红蛋白和β血红蛋白两种类型，替换速率的大幅度提高暗示了蛋白质的功能可能发生了新的转变。

尽管存在着如上诸多不尽人意的地方，分子时钟假说仍然在大多数已经应用的先例中被证明是非常有用和有效的。但是，前提是必须事先估算出你所要研究基因的替换率。

11.1.3 分子进化中的“不确定性理论”

常规的自然选择理论似乎已经很难解释目前这么庞大的生物体的 DNA 功能的多态性了。我们将在本书的后1/3对多态性展开详细的介绍。在本书的第18章，我们将会讨论单核苷酸多态性（SNPs）——一种较为普遍的引起 DNA 功能多态性的现象。

框 11-1

下面是一个有关分子时钟是如何被实际应用的例子，首先，对所研究或给定的蛋白质或者 DNA 的片段的具体分化时间必须能够通过类似于考察化石数据的方法等古生物学方法来估算得到。血红蛋白 C 链的非同义替换率为 0.56×10^{-9}/(位点·年)(由 Gravr 和 Li 于 2000 年提出)。大鼠和人的 α 球蛋白的非同义替换率差异为每个氨基酸位点 0.093。其中同义替换对编码的蛋白质并没有改变，只是改变了其密码子序列而已。例如，密码子 CCU、CCC、CCA 和 CCG 都是特异性地对应脯氨酸的，所以大多数在密码子第三个位点上发生的替换事件都是同义的。而非同义替换将改变其编码的氨基酸种类。例如密码子 ATG 中每个核苷酸的改变都将不能翻译得到原本与之对应的甲硫氨酸。对于这样的密码子，统计其同义替换率就没有任何意义了，因为它根本就不存在同义替换的位点。

核苷酸的替换率 r 是用发生替换的核苷酸数目差值去除以 2 倍的同源序列分化所需要的时间 ($2T$) 得到的，T 表示两个序列分化所需要经历的时间。其中大鼠和人类两物种的分化时间可以被估算出来(参看 2000 年 Graur 和 Li 的原始数据)：

$$r=\frac{k}{2T} \tag{11-3}$$

$$T=\frac{0.093}{(2)(0.56\times10^{-9})}=8.3\times10^{7}\text{(年之前)} \tag{11-4}$$

达尔文（Darwin）的进化论暗示了这样的现象：在表型水平上，那些对于物种继续存活有利的特征将被自然选择保留下来。在分子水平上，一个较为常规的进化观点认为阳性选择对 DNA 序列的多态性的形成起一定的作用。例如，一个原本编码一种酶的基因可能会发生自我扩增（详见第 15 章和第 16 章），随后的核苷酸的改变可能导致其中某个扩增后的基因编码一种具有全新生物学功能的酶，这个新功能对其生存有利并将被自然选择保留下来。

Motoo Kimura 于 1968 年和 1983 年两次提出了用来在 DNA 水平上解释进化的一种特殊的模型。根据他的不确定理论，从分子水平上来看，引起进化改变（或者称为变异性）的主要原因是有关突变等位基因随机漂移的选择是不确定的或者接近于不可确定的。这些突变等位基因的命运是由随机的遗传漂移事件决定的。在这个模型下，达尔文进化论中的阳性选择只起到了非常有限的作用。例如，胰岛素分子中 C 多肽［见图 11-1(b)］的离散性可能就是因为遵从了不可确定的突变率，所以显得相对较为迅速，而 A 链和 B 链因为受到了自然选择的约束所以进化得较为缓慢。

11.1.4　分子系统演化的目标

所有生命形式都共享一个起源，它们都是生命树上的某一部分。超过 99%的曾经在自然界存活的物种都已经灭绝了（Wilson 于 1992 年提出这一观点）。现存物种中，具有较近缘关系的生物体比亲缘关系较远的生物体更有可能是从同一个近代祖先进化而来的。从理论上来说，应该存在着一棵精确描述物种间进化关系的生命树。系统演化的一个目标就是用于推导出那棵包含所有生命形式的正确的进化树。历史上，早期的系统发生演化是根据一些较为简单的可被观察到的特征来分析的，如翅膀或脊索的出现与否。近来，对系统演化的分析也开始着眼于基因和蛋白质家族的分子序列数据。系统演化的另一个目标是用于推断或者估算不同生物体从它

> 生命树只是一个通俗易懂的比喻而已，进化理论未必最后一定能得到一棵独立的进化树，事实上，进化是一个突变和选择共同起作用的过程。我们将在第 14 章中学习到，基因还能在物种间发生水平转移，这是生物体获得基因和特征更为复杂的一种方式。

们上一级共同祖先开始分化的具体时间。

生命树主要存在着3个分支：细菌、古细菌和真核生物。我们将在第12章对全生命树做进一步探讨，在该章中我们将对用于评估同源蛋白质家族（或者同源氨基酸序列）的系统发生树做详细的介绍。

真实的生命树是用来描述进化历程中发生的真实历史事件的，但是生成一棵真正的进化树几乎是不可能的。作为替代，我们将生成一棵推论的进化树，用于描述经研究得到的推测的历史事件。这些进化树往往通过研究那些已知的数据并利用一些模型来研究描述一系列进化事件。

在第3章中，我们将两个具有共同祖先的蛋白定义为同源蛋白。你能够运行一个BLAST搜索并且观察到一些蛋白获得较高的分值（和低期望值）并且简单地认为这些蛋白可能具有相关的功能。但是，对进化信息中的旁系同源（orthologs）和直系同源（paralogs）信息还是要给予一定的关注。我们应用了大量的方法来研究蛋白质之间的关系：用Dayhoff的打分矩阵来进行成对的比对分析（详见第3章）、BLAST搜索（详见第4章和第5章）、蛋白质折叠子关联度的研究（详见第8章和第9章）和多重序列比对（详见第10章）。所有的这些方法都是来源于进化的模型，通过统计分子序列水平的差异性和相似性来确定的。

Theodosius Dobzhansky在1973年曾经提出过："在生物学研究中除了根据进化研究出来的结果几乎没有其他有道理的结论。"的偏激理论。

① Dayhoffet等在1978年通过一个清晰的进化谱系来提出他的打分矩阵："一个可接受的蛋白质点突变是指其中一个氨基酸位点被替换后，它仍然能被自然选择保留下来。这种结果的产生主要通过两个显著的过程：首先，部分基因模板上产生了一个核苷酸的点突变，并产生了包含点突变氨基酸的相应蛋白质；其次，物种接受甚至更加倾向去选择这种对生存更为有利的突变形式。被接受的含有点突变氨基酸的蛋白质必须与原蛋白具有相似的生物学功能。实验观察得知，那些在化学和物理性质方面都保持高度相似性的氨基酸经常发生互换。" Dayhoffet等研究两种蛋白质的氨基酸序列时，并不是将两者进行序列比较而是将它们各自与它们在系统发生树上推得的共同祖先进行比较研究。

② Feng和Doolittle在1987年利用Needleman和Wunsch的成对比对的方法发展出了一套蛋白质多重序列比对的方法，并且据此得到的蛋白质相互关系来重建进化树。那些序列被假定倾向于来源于一个共同的祖先，进化树是根据多重序列比对所得到的各种不同的矩阵直接重建而来的。对于这个方法，输入更多近代才进化分离的序列比输入远古时代（即进化分离）所得到的结果可信度要高得多。

③ 在下文有关蛋白质家族的描述中，我们会用一种包含进化距离信息的多重序列比对的Pfam JalView工具来分析任何一个Pfam家族，并将其用进化树的形式描述出来（见图10-17）。

在本章中，我们将利用蛋白质（或者DNA和RNA）序列的多重比对信息来生成系统发生树。这些系统发生树为研究者提供了一种分子序列进化历史研究的可视化手段。

11.2 分子系统发生：系统发生树相关专用名词

分子系统发生是一门利用分子生物学技术来研究不同物种或者分子间的进化关系的学科。更多其他生物技术目前也被频繁地应用在进化研究上，包括形态学、解剖学、古生物学和生理学。我们将利用分子序列数据来研究系统发生树，用专业术语来描述进化树的特征。

从任何系统发生树中都能得到两类关键的信息：树的拓扑学性质和分支长度。下面将对系统发生树的一系列特征做一必要介绍。

一棵树的拓扑结构决定了其上蛋白质（或者其他对象）之间的关系。例如，树的拓扑结构显示了两个同源蛋白质序列的共同祖先。树的长度某些时候（但不是始终）反映树中目标的相关度。

让我们首先来定义树的主要部分和主要类型的树。一棵系统进化树是一个由边和点组成的图［图 11-4(a)］。每个分支（也叫做边）都连接两个节点。节点是用来描述分类单元的；节点（node，这个概念来源于拉丁语“knot”）是两个或多个分支的交叉点或终点。对于我们，分类方法可能用蛋白质序列作为典型特征。树的外部节点或者叶节点上的操作单元（OTU）是一个具有显著特点的分类，例如对我们所分析树的可接受的核苷酸或者蛋白质序列的差异性。

在我们讨论的树中，我们假定所有的原始数据都包括 DNA、RNA 或者蛋白质序列数据。这些数据都是通过多重序列比对得到的。

参考图 11-4(b)、图 11-4(c)，每棵树都包含了 5 个操作分类单元（OTUs）(分别标以 A、B、C、D 和 E)。另外，还有在 F、G、H 和 I 位置的内部节点。每一个内部节点都代表操作分类单元（OTUs）所推测得到的祖先。想像一下这棵进化树代表了 5 种 lipocalins（一个运输小疏水分子的蛋白家族，这种蛋白可以调节泪膜的表面张力），其中 A 和 B 分别代表人和老鼠的 RBP。连接 A 和 B 节点的内部节点，代表了一个 80MYA（百万年前）在灵长类和啮齿类动物演化之前出现的古生物体内的祖先序列。分支的情况定义了树的拓扑结构，它表示了不同祖先分类之间的相互关系。在有些进化树上，分支的长度表示了该分支上氨基酸的改变数目。一棵树的分支也叫做边。在图 11-4 的树中，每个连接到操作分类单元（OTUs）上的边叫做外部分支（或者叫做外围分支）。那些连接到 F、G、H 和 I 节点上的分支叫做内部的分支。

在图 11-4(b) 的例子中，树是不标示权重的。这就表示其分支和氨基酸改变的数目不成比例。这种进化树（也称作进化分支图，cladogram）的表现形式有其独特的优点，即能够将操作分类单元整齐地对齐在一条垂直的纵队上。这个优点在该树存在很多操作分类单元（OTUs）时特别有用。而且，这种形式的树还可以按照推测得到的进化分支事件分化的大致时间尺度来编排那些节点的相对位置。

在图 11-4(c) 的树中，用与图 11-4(b) 相同的原始数据来生成另一棵树，所不同的是这棵树的分支长度是标示加权长度的。因此，每个分支的长度都与不同序列中氨基酸（或者核苷酸）改变的数目成比例。这种形式的进化树（称作种族分支图，phylogram）的优点在于其能够非常清楚地可视化地反映书中不同蛋白质的相关性。

当一个内部节点只有两个直接的线性后代（或称分支），那么这个节点将是二分叉的。而二分叉的树通常称为二叉树或者叉状分支树，其中任何一个分支都直接分裂成两个子分支。如果一棵树存在着包含两个以上直接后代的节点（例如图 11-5，ABC 分支上的内部节点）就称为多分支树。

多分叉也称作三分以上的状态。多分叉树也可以定义为非二叉树。

一个进化分支包含一个共同祖先和它演化得到的一切分类。一个进化分支也称为单源组。在我们的上下文中，进化分支指的是进化树中形成一组分类的蛋白质。以图 11-4(b)、图 11-4(c) 的任何一棵树为例，C、D 和 H 共同形成一个进化分支，但是 B 不属于这个分支中的一分子。一个大的进化分支包括 C、D、H、A、B、F 和 G。但是，操作分类单元

（OUT）中标记为E的节点不是这个大分支中的一部分。分类组ABF和分类组CDH共享一个共同的祖先G，称为一对平行种族。

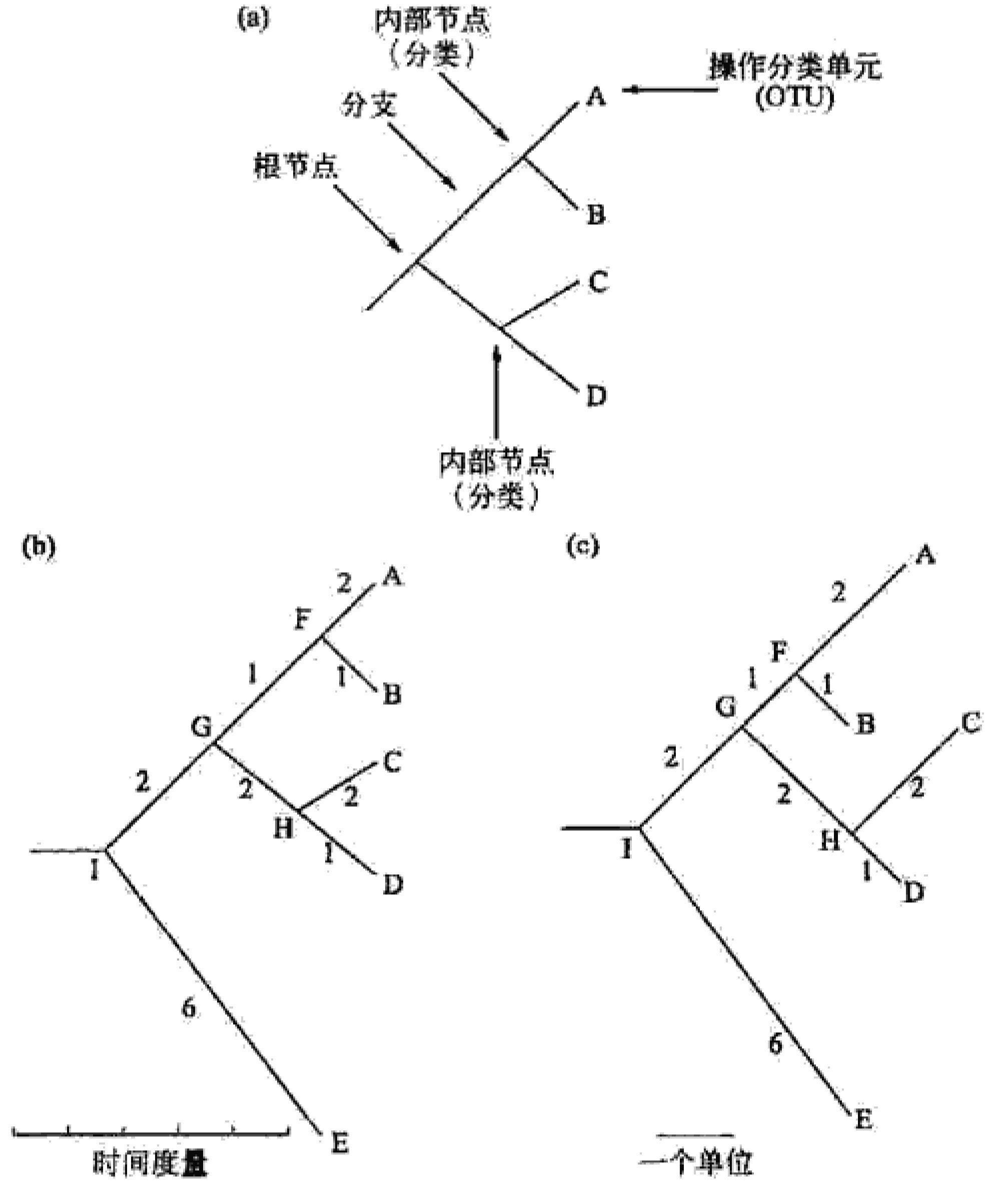


图 11-4 （a）系统发生树包括点和边。一个节点可能是一个外部节点，或者中间节点或者是树的一个根节点（根节点将在下文做详细介绍）。每条边都有两个节点。节点代表分类或者一个分类单元；一个包含可观察的特征（例如现有的蛋白质序列或者形态学上特征的分类单元）一般被称为操作上的分类单元（OTUs）。系统发生树有标示尺度的（b）和不标示尺度的（c）两种。在一棵不标示尺度的树中，树分支的长度并不与氨基酸（或者核苷酸）的改变数目成比例。例如，FA分支上的点（两个单位）和FB分支（一个单位）在（b）中有非常明显的长度区别。在图中的操作分类单元（ABCDE）都是整齐排列在树的末端成一列的。x轴表示时间尺度（例如每个分类单元中的单位为数百万年）。在带标示长度的树（b）中，分支的长度是与替换的数目成比例的。根据这个拓扑性质我们能更为简单地观察树中蛋白质（或者基因）之间的相互关系

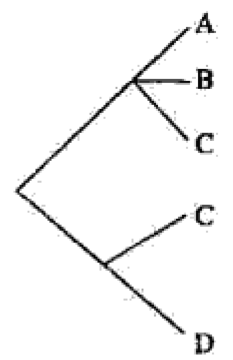


图 11-5 一棵含有3个或3个以上节点或分支的进化树一般称为多分支树或多叉树（参看引导至ABC分类的节点）。当没有足够的可获得信息来确定该进化树是一棵二叉树时研究者通常构造多分叉树

11.2.1 树根

一棵系统进化树如果是有根树，则代表所有的序列都共有一个常见的祖先。但是大多数情况下树的根节点目前还是不确定的，一些构建进化树的算法甚至不为树根预留位置。相对于有根树的另一个选择就是构建无根树。无根树将用于确定 OTUs（操作分类单元）之间的相互关系。但是，它并没有定义完整的进化路径或者推测其共同的祖先。图 11-6 展示了一棵带有 5 个 OTUs（操作分类单元）的二叉树，有无根的树［图 11-6(a)］也有有根的树［图 11-6(b)］。OTUs（操作分类单元，现存的分类，叶节点）用 1～5 来标记。一些 OTUs（操作分类单元）的交换（调换）的改变对树的拓扑结构的改变没有任何作用，例如任一棵树中的 4 节点和 5 节点。但是，某些节点是不能被交换的，例如 1 节点和 2 节点。值得注意的是，无根树进化时间的方向有时难以确定。

一个确定树根的主要方式是确定外围分类的进化关系。在图 11-6(c) 中，将 1～5 的序列假想为哺乳动物 RBP 的直系同源分子并且这些序列是从一同源细菌或者无脊椎动物蛋白（OTU 6）进化而来的。我们现在已经非常清楚地知道这种无脊椎动物序列是由一个在所有其他的 OTUs（操作分类单元）之前出现的序列衍化而来的，所以它可以被用来定义树根的位置。

通过确定最近的外围分类我们可以选择一个 OTU（操作分类单元）来确定树根的位置。第二种方式是通过中间生根法来确定树根。这里，最长的进化树分支已经被确定了［例如图 11-6(a) 中内部节点 7 和 8 之间的分支］，通常情况下最长的分支一般都被假定为最合理的根节点所在位置。

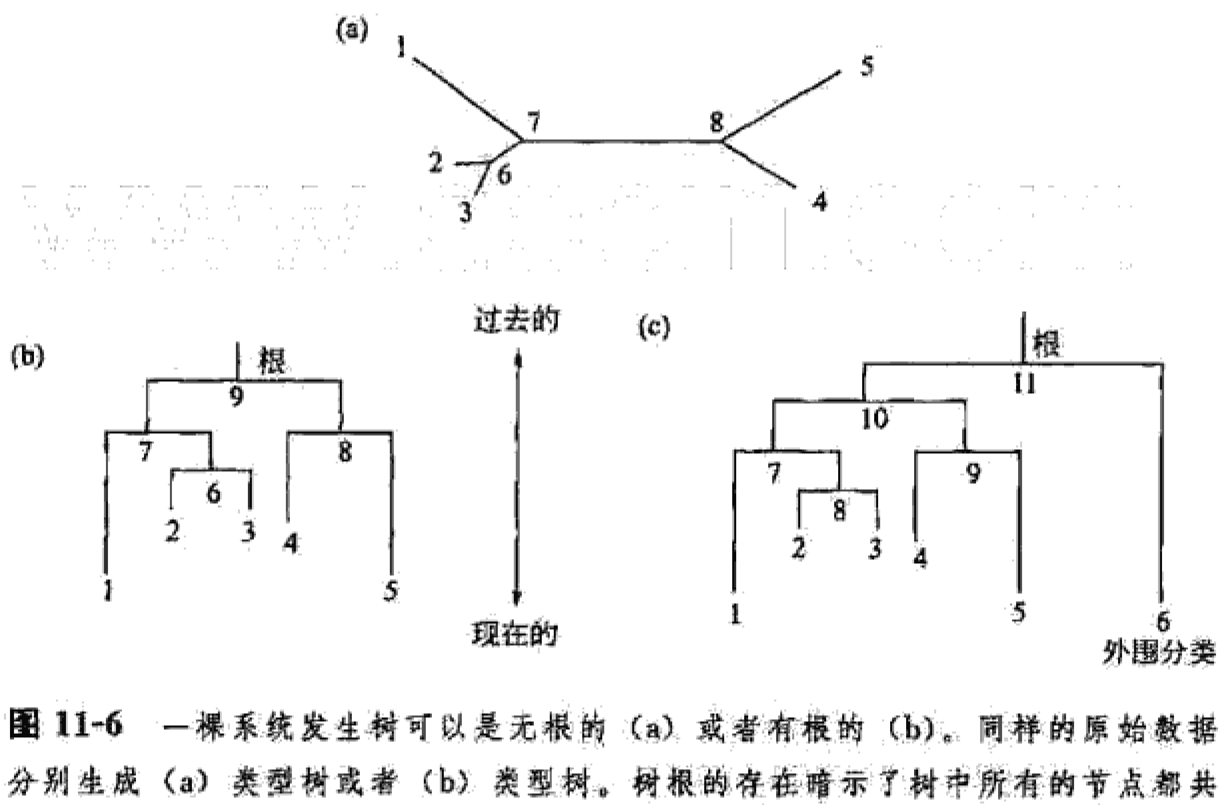


图 11-6 一棵系统发生树可以是无根的 (a) 或者有根的 (b)。同样的原始数据分别生成 (a) 类型树或者 (b) 类型树。树根的存在暗示了树中所有的节点都共享了同一个最原始的祖先。但对于那些祖先信息未知或缺失的数据，无根树的解释更为适当。给进化树确定树根通常有两种常见的方式。正中央根源于树中最长的一个分支［图 (b) 中，此分支连接 7 和 8 两个节点］上中间点可能是添加根节点的最佳位置。(c) 一个系统遗传距离上相近的生物体单独分类被添加到数据集中（分类 6），并且在新建的树中一般定义为一个外围分类 (outgroup)。例如，一个无脊椎 lipocalin（一个运输小疏水分子的蛋白家族，这种蛋白可以调节泪膜的表面张力）可能就是哺乳动物一系列 lipocalin 中的一个外围分类 (outgroup)

11.2.2 枚举树

对于一打以上的蛋白质序列来说，可能的进化的形式是非常多的。对于你自己建立的任何一棵树，知道可能存在的树的数目有时非常重要。其中只有一棵“正确”的树能够代表蛋白质分子序列（或者是物种之间）进化的途径。

必须清楚的是：细菌的 lipocalin（一个运输小疏水分子的蛋白）可能在几十亿年前或者几十亿年前的时候就已经出现。无论它在地球上存在的时间多长，它的进化速率也不一定比其他任何一个物种的 lipocalin 的快或慢。

框 11-2 有根树和无根树的数目

二叉无根树（N_U）的 OTUs 数目 n（$n \geqslant 3$）是 CavalliStorza 和 Edwards 在 1967 年提出的：

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

二叉有根树（N_R）的操作分类单元（OTUs）的个数 n（$n \geqslant 2$）是：

$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

例如，对于 4 个 OTUs，N_R 等于 $(8-3)!/[(2^2)\times(2)!]=5!/8=15$。

有根和无根的树可能的数目（大于 50 个 OTUs）类似于下面表格中情况。这些值是通过 Matlab 软件（MathWorks）计算得来的。

OTUs 数量	有根树的数目	无根树的数目
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10395	945
8	135135	10395
9	2027025	135135
10	34489707	2027025
15	213458046676875	8×10^{12}
20	8×10^{21}	2×10^{20}
50	2.8×10^{76}	3×10^{74}

想像一下，仅一小部分分类对应的可能的树的数量非常大，宇宙中这个数目大约是 10^{77}。

构造进化树算法的应用对于确定潜在的树的数目是十分有用的。有根树和无根树存在的可能数量在框 11-2 里有详细描述。对于 2 个 OTUs，只存在一棵可能的树。对于 3 个，可能构造出 1 棵五根树或者 3 棵不同的有根树（图 11-7）。对于 4 个分类，可能构造出的树上升到 3 棵无根树或者 15 棵有根树（图 11-8）。一个重要的实际限制是这种枚举法的应用只能局限在 10 个序列，此时可推测出来的有根树有 3400 万棵，无根树有 200 万棵。所以就非常有必要利用启发式算法的方法来评估这么多树中最有鲁棒性的树了。与之形成对比的是，对于 10 个或者更少的分类，利用标准的台式电脑就能对所有评估认为可能候选的树进行穷举搜索。

某些进化的项目包括上千种分类的进化树。参见“深绿植物计划”（http://ucjeps.berkeley.edu/bryolab/greenplantpage.html）。核糖体数据库（http://rdp.cme.msu.edu/html）包括超过 50000 条对齐后的序列分析。对于典型的分析，你可以分析一堆的分类。如果要利用 Pram（版本 9.0），你能够利用 58 个可以利用的蛋白质序列作为种子序列进行比对，或者对 396 个可获得的蛋白质做全序列比对。

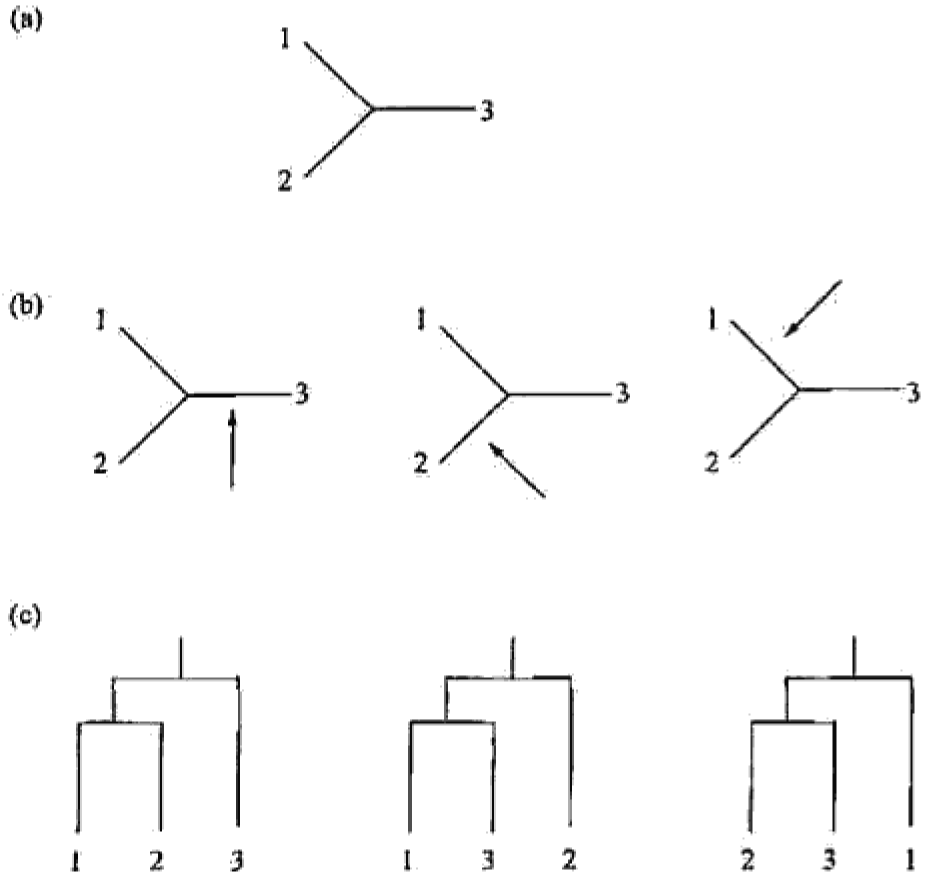


图 11-7 (a) 对于只有 3 个操作分类单元的进化树（例如 3 个对齐后的蛋白质序列 1～3），与之对应的只有一棵无根树；(b) 树上的任何一条边都可以被用作树根的候选对象；(c) 3 个相应的可能存在的有根树

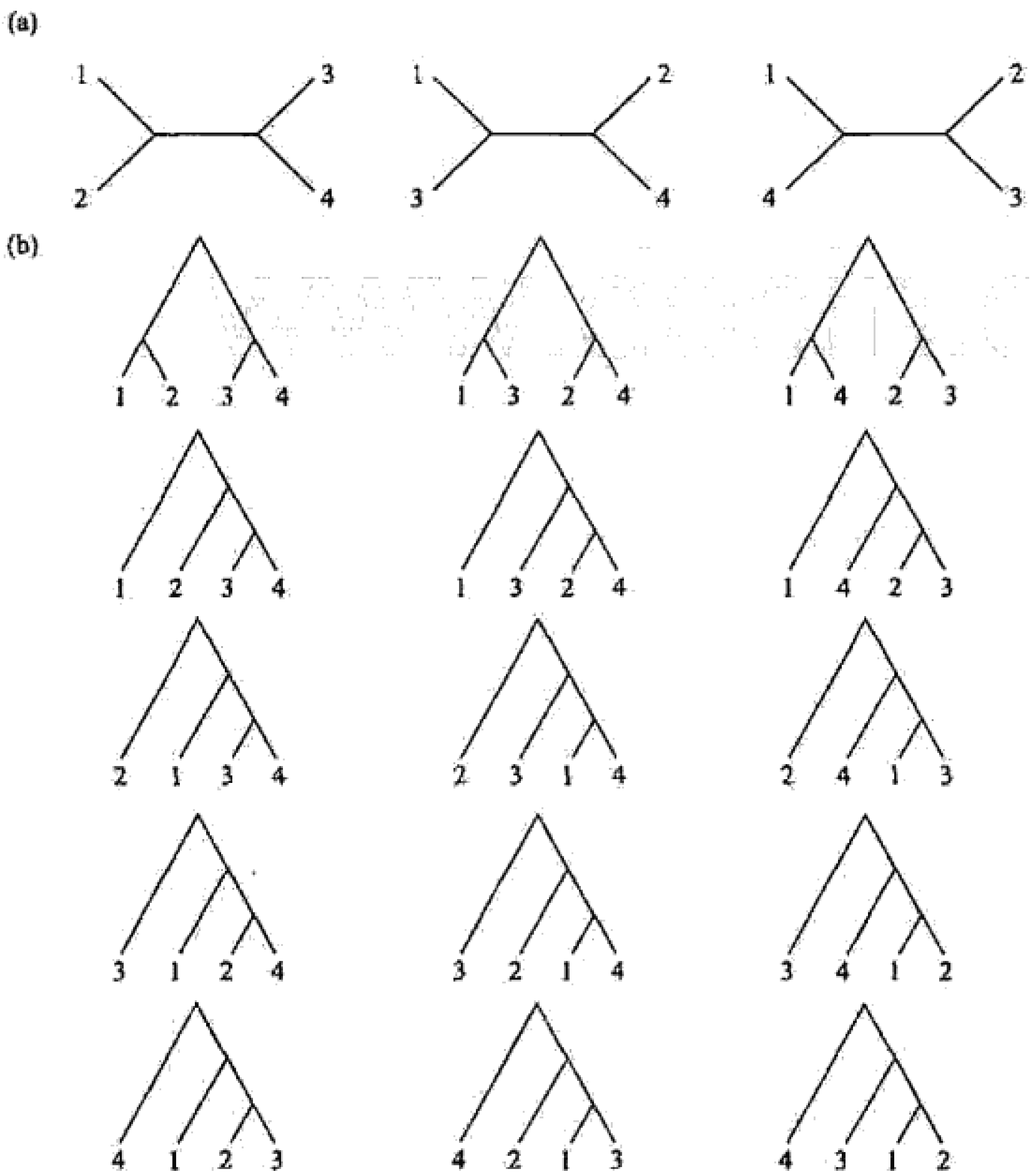


图 11-8 对于 4 个操作单元分类（例如 4 个对齐的蛋白质序列 1～4），一共有 (a) 3 棵可能的无根树和 (b) 15 个可能存在的有根树。只有一棵树是在拓扑学结构上正确描述这些序列发展的进化历程的

11.2.3 物种树和基因/蛋白质树

物种进化和蛋白质（和基因）进化。有关蛋白质进化分析的复杂度在于两个物种分化时间上的复杂性。特殊的例如：两个新物种都是从一个单独的祖先物种进化而来的，这种特例中，一般新物种都出现在祖先形成生殖隔离以后（图 11-9）。在一个物种树中，一个内部节点代表一个物种形成的过程。在一棵基因树（或者蛋白质树）中，一个内部节点代表一个祖先基因分化为两个新的独特的基因序列的事件。

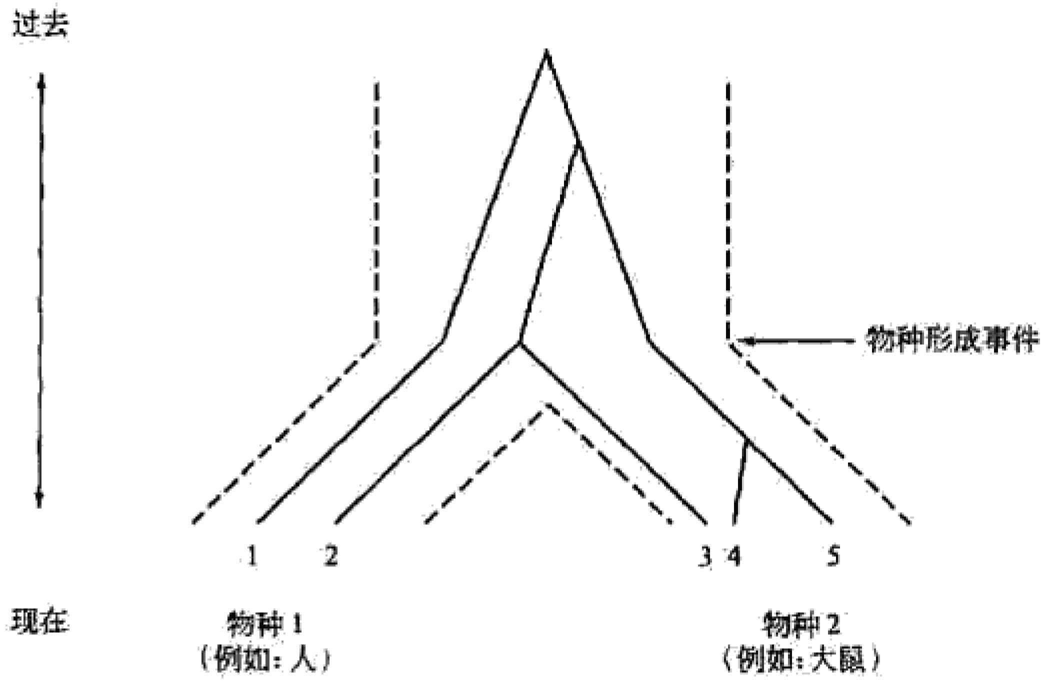


图 11-9 一棵物种树和一棵蛋白质（或者基因）系统发生树可能有较为复杂的相互关系。一个物种形成事件（例如产生现代人类和啮齿动物的分化）能够被估算到一个特殊的时间（例如 80MYA）。当物种形成发生时，物种之间发生生殖隔离。这个事件用带点的线来表示（见水平箭头）。如果存在基因的自我复制发生在物种形成事件产生之前或者之后的一组同源蛋白组合的系统发生分析将会较为复杂。现存物种的蛋白质的（OTUs）序列可做本质上的系统发生分析。重建蛋白质家族的历史和重建每个物种的历史一样重要。在上面的例子上，一共有 2 种人类的直系同源子和 3 种大鼠的直系同源子。蛋白质 1 和蛋白质 5 分化发生在两个物种分化之前。蛋白质 2 和蛋白质 3 的分化与物种的分化同时发生。蛋白质 4 和蛋白质 5 在最近发生分化，即在物种分化之后发生分化。理解这些有利于重建物种树和蛋白质（或者基因）树。Nei 在 1987 年和 Graur 在 2000 年分别对之进行过修改

在一个遗传上呈现多态性的种群中，基因的自我复制事件可能发生在物种形成之前或者之后。一个蛋白质（或者基因）树与物种树主要在两个方面存在差异性（Graur 和 Li，2000）：①来自两个物种的两个基因的分化可能发生在物种形成事件之前。这就可能导致对系统发生树中最长的分支的分析出现过估计的现象。②基因树的拓扑结构可能和物种树存在差异。一个特殊的例子就是，通过研究一棵基因树去重建一棵物种树可能是比较困难的。分子时钟一般被用于估算基因树上的基因发生分化的时间，但是这个时间不能假定为也是物种形成的时间。

通过利用一个单独的蛋白（或者基因）重建一棵系统发生树有时候能够得到很复杂的结果。因为这个原因，很多研究者利用大量独特性质的蛋白质（或者基因）家族来构建进化树，以此来估算不同物种之间的相互关系。另一个可以被接受的策略是产生更多的连接蛋白（或者DNA）的序列。例如，Baldauf 等在 2000 年利用 4 个连接蛋白序列来建立一个真核细胞全面的系统进化树（图 16-1）。

启发式的算法能够被用于研究所有可能树的子集，去掉大部分在拓扑学上不合理不可用的进化树。利用启发式的算法（见框 2-2），可以构建含有数百个蛋白（或者 DNA）序列的系统发生树。一个关于启发式算法是如何起作用的例子是：想像一下在数据集中用算法搜索总分支长度最短的进化树（如最节俭树）。这种搜索方法并不遍历评估所有可能的进化树，取而代之的是评估那一系列重排的拓扑结构。一旦一棵进化树获得一个比较特殊的分数，这个算法就能够设这个特殊的分数为上限并且抛弃所有不可能比它短的拓扑结构的进化树。

11.3 系统发生分析的 4 个步骤

分子系统发生分析能够被分为 4 个主要的步骤：①选择可供分析的序列；②同源蛋白或者核酸序列的多重序列比对；③构建系统发生树；④进化树的评估。

11.3.1 第一步：利用 DNA、RNA 或蛋白质序列来进行分子系统发生分析

当利用分子序列数据来构建生成系统发生树时，你可以用 DNA、RNA 或者蛋白质的序列。在某一个假定的环节里，你可能需要评估一组分子间的相互关系，例如评估一组 lipocalins 之间的进化关系。至于到底是选择以蛋白质还是 DNA 为研究对象取决于所要研究的对象和想要解决的问题。在某些例子中，对蛋白质的研究是更为可取的；你可能倾向于研究蛋白质的多重序列比对，或者是蛋白质相对于 DNA 而言更低的氨基酸替换率可能使其更加适用于比较广泛的分化的物种。在另外一些例子中，研究 DNA 能够比研究蛋白质获得更为丰富的信息量，其中的原因有多方面。

① 编码蛋白质的那部分 DNA 可以发生同义或者非同义的替换事件。例如下面 5′末端（蛋白质的氨基末端）和 3′末端（蛋白质的羧基末端）的 3 种 RBPs 序列的比对分析（详见图 11-10 的箭头 11～14 和箭头 21～34）。

② 比较非同义替换率（$\hat{d}_N$）和同义替换率（$\hat{d}_S$）可能会显示出物种是发生阳性选择还是阴性选择的证据。如果 $\hat{d}_S$ 大于 $\hat{d}_N$，这就暗示 DNA 序列可能发生了阴性选择或者净化选择。阴性选择是对相应的氨基酸序列改变的限制；这种选择发生在蛋白质的结构和/或者功能是临界的和不能发生替换的时候。当 $\hat{d}_N$ 大于 $\hat{d}_S$ 时暗示了阳性选择的发生。阳性选择的一个例子是基因自我复制并且在自然选择的压力下进化出新的功能。

现在有一种较被接受的观点是突变事件发生的数目要比我们利用 PAM 矩阵的方法（详见第 3 章）能够直接观察到的多得多。这里，我们观察到两种长度在 100 个氨基酸量级的蛋白质分享了 50% 的氨基酸相似性，并且保持了大约平均 80 个的改变数目（图 3-19）。

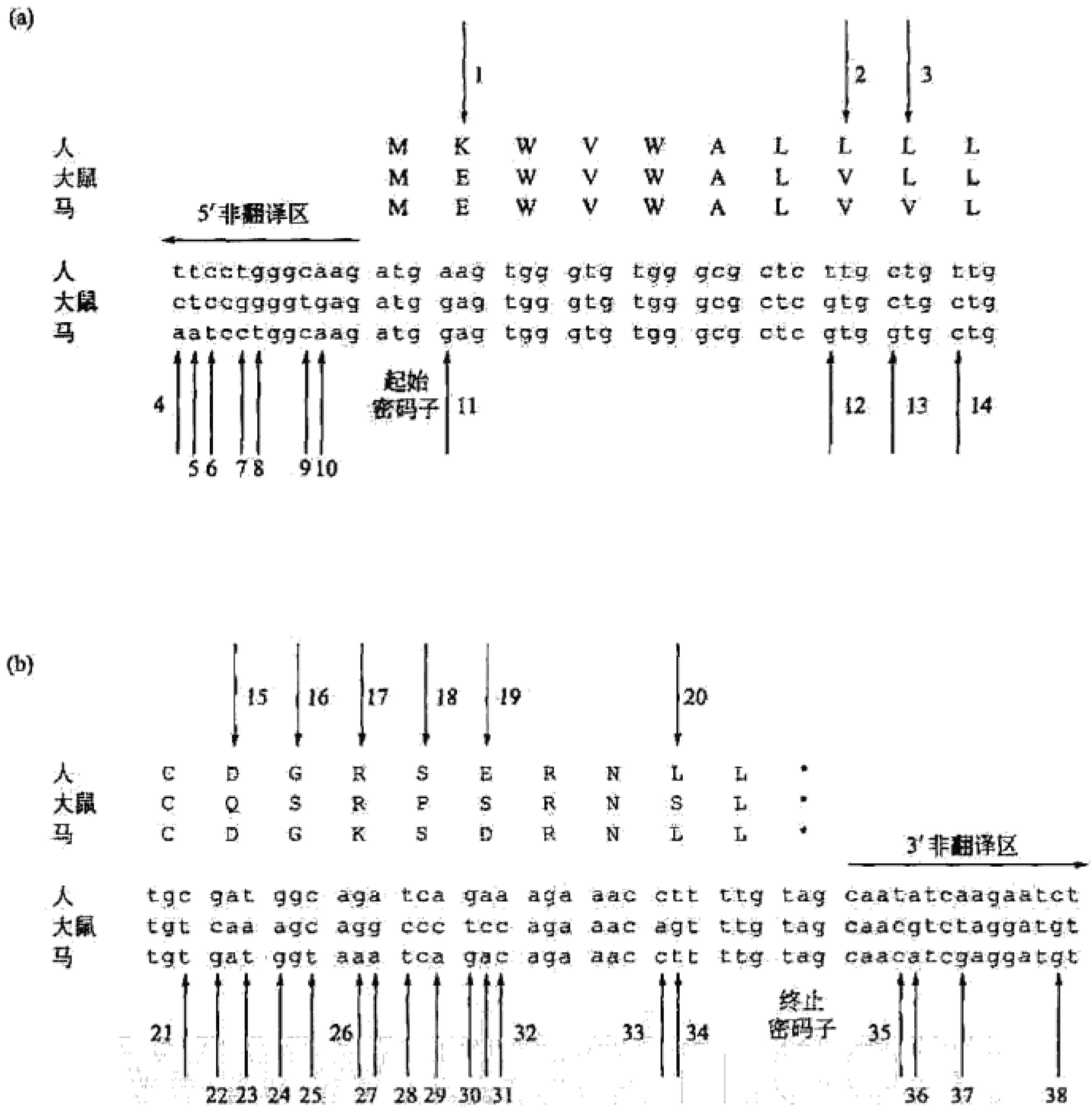


图 11-10 系统发生树可以用 DNA、RNA 或者蛋白质序列数据来构建。通常情况下，对于系统发生分析，DNA 序列所包含的信息量要比蛋白质序列多。一个很好的事例就是，我们将 3 个物种的松香油结合蛋白的序列从 DNA 的 5′端开始对齐［和（a）中蛋白质相应的氨基末端对齐］一直到 DNA 的 3′末端［与（b）中相应的蛋白质的羧基末端对齐］。在 5′端和 3′端的非翻译区域，不编码任何蛋白质，因此它们维持特殊氨基酸残基所需面临的选择压力也将小得多（某些调控元件是高度保守的）。这里，在 5′端共有 7/12 核苷酸位点是不保守的（参见箭头 4～10）在 3′末端不保守的核苷酸位点相对较少（参见箭头 35～38）。在蛋白质编码区域，前 10 个氨基酸残基位点中共有 3 个不保守位点（箭头 1～3），后 10 个氨基酸残基位点中共有 6 个不保守位点（箭头 15～20）。这些不保守位点的存在可能是演绎系统发生的有用信息。然而，与观察到的数目相比，事实上存在着更为大量的信息核苷酸的改变，限制了我们更多地关注编码的区域：作为非同义替换附加物（引起氨基酸的改变：参见图中箭头 11～13、箭头 22～24、箭头 26、箭头 28、箭头 30～34），这里只存在少量的同义改变（核苷酸的改变不引起氨基酸的改变；参见图中箭头 14、箭头 21、箭头 25、箭头 27、箭头 29）。图中 RBP 的序列数据来源于人类的基因组（GenBank 序列号：NM_006734）、老鼠的基因组（GenBank 序列号：XM_129259），以及马的基因组（GenBank 序列号：U21208）

③ DNA 中发生的替换（包括直接在比对中观察到的），例如单核苷酸的替换、连续的替换以及巧合型替换，在图 11-11(a) 和图 11-11(c) 中有详细的描述。通过分析两种序列并且参考其祖先序列［图 11-11(b)］，这可能可以推论出大量在两个（或者更多）直接序列比对中得不到的有关突变的有用信息。这些突变的过程包括平行置换、聚集置换和回复置换［图 11-11(c)］。

共源性状（synapomorphy）一般被定义为一个被几个分类共享的特征。同形性状（homoplasy）被定义为一个独立出现的特征（例如，伴随在聚集替换和回复置换过程中出现的），却是从一共同祖先衍生而来的（但其彼此可能并不一定存在同源关系）。详见 Graur 和 Li 在 2000 年的研究成果。

④ 非编码区（例如在内含子上 5′端和 3′端的基因非翻译区）可能被用于分子系统发生分析（图 11-10）。在大多数例子中，对于这部分氨基酸序列保守性的进化压力非常小，这部分氨基酸区域通常变化性较大。于是，这段氨基酸区域的替换率与中立突变率相等。在其他例子中，存在着巨大数量的氨基酸保守性，可能是因为类似于转录结合因子的绑定位点的调控元件出现的缘故。

⑤ 通过分子系统发生的方法我们还可以研究假基因。根据定义来看，这些基因是不编码功能蛋白的（参见第 12 章）。

⑥ 转换和颠换的速率能够被估算。在某些例子中，转换的发生频率要比颠换的发生频率至少大 10 倍（Moritz 和 Hillis，1996）。

A 发生转换一般对应两个嘌呤（A 变成 G 或者 G 变成 A）或者对应两个嘧啶（C 变成 T 或者 T 变成 C）。A 发生颠换一般在一个嘌呤和一个嘧啶之间（例如：A 变成 C，C 变成 A，G 变成 T 等，一共有 8 种可能的颠换方式）。

虽然 DNA 的分析带给研究者很多优点，但是人们还是倾向于将蛋白质的序列用于系统发生的分析。两种生物体中的进化距离可能是太遥远了，以至于任何一个 DNA 序列相对而言都是饱和的。这是因为在大部分位点上，所有可能的核苷酸改变都将会发生（以偶倍数翻倍），所以导致系统发生信号的丢失。蛋白质一共有 20 种氨基酸来替代只有 4 种核苷酸的 DNA，所以就产生了更强的系统发生信号。对于关系较为接近的序列（例如大鼠和小鼠的 RBP），基于 DNA 序列的系统发生分析比基于蛋白质研究的系统发生分析可能更为合适，因为基于 DNA 序列产生的与两者相关的系统发生分析信号相对更强（图 11-10 和图 11-11）。

国际应用和纯粹化学学会（IUPAC；http://www.iupac.org）定义了大量科学中普遍应用到的符号。4 个核苷酸简写分别是：腺嘌呤（A）、胞嘧啶（C）、鸟嘌呤（G）以及胸腺嘧啶（T）。一些扩充的简写为：对于一个不确定或者未知的核苷酸一般简写为 N，一个未指明是何种嘌呤的核苷酸缩写成 R，另外一个未指明是何种嘧啶的核苷酸缩写成 Y。

无论是选择核苷酸还是氨基酸来做系统发生分析，字母改变所带来的结果还是可以被定义的。一个无序的字母改变是一个核苷酸或者氨基酸一步一步地改变成另一个字母的过程。一个有序的字母改变是一个必须经过 1 个或者更多个中间步骤后才能变成不同字母的过程。部分排序字母在开始值和终止值之间有一系列可变的不确定数目的中间状态过程。核苷酸是

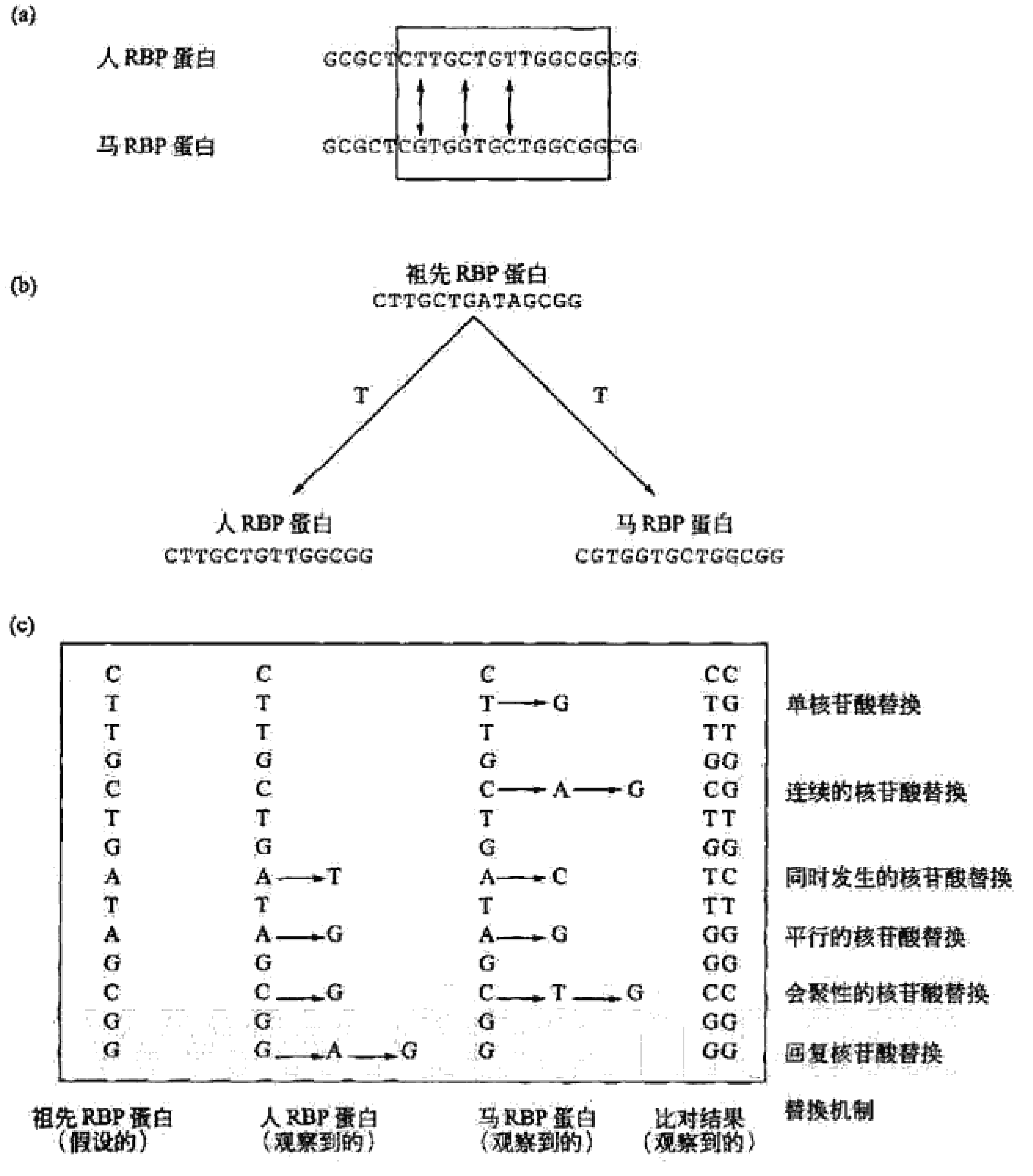


图 11-11 发生在序列中的突变形式有很多种。(a) 人和马的 RBP 编码序列的一部分，包括 3 个可观察到的错配。事实上在这个区域内有远比 3 个更多的突变事件发生。(b) 假设存在一个人和马的 RBP 编码序列的祖先序列在时间 T 后分化为人和马的 RBP 序列。(c) 将人和马的 RBP 编码序列与假设的原始序列对比可以发现很多的突变机制。单核苷酸替换；有序替换和一致替换都能够通过被观察到的突变所统计得到。平行的、会聚的以及回复替换事件都能在不被观察到的情况下发生。在这个事例中，当 12 个突变事件真正发生时我们可观察到的突变事件只有 3 个。(a)、(c) 的数据来源于 Graur 和 Li2000 的研究成果授权使用

非排序字母：任何一种核苷酸都可以在一步改变中变成另外一种核苷酸［图 11-12(a)］。氨基酸是部分有序的。如果观察遗传密码，你将会发现某些氨基酸的改变可以通过一步由一个核苷酸的替换来完成，但是其他某些氨基酸的改变需要 2 个甚至 3 个核苷酸的突变［图 11-12(b)］。

我们将会在第 15 章讨论一个真菌类基因组全序列自我复制的现象。此类现象的证据是通过用 Blastp 搜索比对全啤酒酵母（*Saccharomyces cerevisiae*）蛋白质组的方法来得到，结果显示在大量序列中存在许多保守的序列区域（图 15-9）。这里，Blastp 搜索比对方法可能对不同染色体之间的同源性表示不够敏感。

(a)

	A	C	T	G
A	0	1	1	1
C	1	0	1	1
T	1	1	0	1
G	1	1	1	0

(b)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	2	1	1	2	1	2	2	2	2	2	2	1	2	2	1	1	1	2	2
C		0	2	3	1	1	2	2	3	2	3	2	2	3	1	1	2	2	1	1
D			0	1	2	1	1	2	2	2	3	1	2	2	2	2	2	1	3	1
E				0	3	1	2	2	1	2	2	2	2	1	2	2	2	1	2	2
F					0	2	2	1	3	1	2	2	2	3	2	1	2	1	2	1
G						0	2	2	2	2	2	2	2	2	1	1	2	1	1	2
H							0	2	2	1	3	1	1	1	1	2	2	2	3	1
I								0	1	1	1	1	2	2	1	1	1	1	3	2
K									0	2	1	1	2	1	1	2	1	2	2	2
L										0	1	2	1	1	1	1	2	1	1	2
M											0	2	2	2	1	2	1	1	2	3
N												0	2	2	2	1	1	2	3	1
P													0	1	1	1	1	2	2	2
Q														0	1	2	2	2	2	2
R															0	1	1	2	1	2
S																0	1	2	1	1
T																	0	2	2	2
V																		0	2	2
W																			0	2
Y																				0

图 11-12　(a) 核苷酸或者 (b) 氨基酸对应的步幅矩阵描述了从一个字母改变到另外一个字母所需要经历的步骤数目统计值。对于氨基酸，通过 1～3 个核苷酸的突变步骤就可以将一个残基改变为另外一个残基。数据来源于 Graur 和 Li (2000)，被授权使用

11.3.2　第二步：多重序列比对

多重序列比对（参见第 10 章）是系统发生关系分析过程中关键的一步。在很多例子中，排列成队的核苷酸或者氨基酸残基暗示了它们可能共享了同一个祖先（参见下文）。如果你错误地比对了一组序列，你仍然可能生成一棵进化树。但是，这样建立的进化树似乎没有任何生物学意义。如果你对包含非同源的序列建立多重序列比对组，非同源的那组序列仍然会被整合到系统发生树结构中去。

为系统发生分析准备多重序列比对组，需要注意几个比较重要的方面。让我们通过考察下文中 13 个特殊的直系同源的 RBPs 例子来介绍其中的注意事项。我们利用这些蛋白质在图 3-2 中建立一棵系统发生树。产生这棵系统发生树多重序列比对的方法在图 11-13 中给出了部分说明。其中有很多值得引起我们注意的特征。

① 仔细检查所有参加多重比对的序列，以保证其中所有的序列都是同源的。有时我们可以通过观察得到一序列与其他的序列距离很大的情况来推断其与其他序列可能是非同源的关系。你可以通过成对比对和 BLAST 搜索或者单独使用其中的某一种来更进一步验证你的推测结果（考察其产生的期望值是否具有显著性）。如果检测发现这个序列显然是非同源的序列，我们就应该将其从多重序列比对中移除。

```
                                               1            2   3
              1                                                         50
鱼类   ccrbp  MLRLCIALCV LATCWAQDFL ESNTTVKQDC ALGTCWAQDC LVSNITVKQD
       drrbp  MLRLCIAVCV LA........ .......... ...TCWAQDC QVSNFAVQQD
       omrbp  ---------- ---------- ---------- -------SDC QVSNIQVMQN
       sarbp  ---------- --------MT RMLRYVVALC LLAVSWAQDC QVANIQVMQN
其他   mmrbp  ---------- -----MEWVW .ALVLLAA.. LGGGSAERDC RVSSFRVKEN
动物   rnrbp  ---------- -----MEWVW .ALVLLAA.. LGGGSAERDC RVSSFRVKEN
       btrbp  ---------- ---------- ---------- ------ERDC RVSSFRVKEN
       ssrbp  ---------- -----MEWVW .ALVLLAA.. LGSAQAERDC RVSSFRVKEN
       ecrbp  ---------- -----MEWVW .ALVVLAA.. LGSAGAERDC RVSSFRVKEN
       hsrbp  ---------- -----MKWVW .ALLLLAA.. W..AAAERDC RVSSFRVKEN
       ocrbp  ---------- -----MEWVW .ALVLLAA.. LGSGRGERDC RVSSFRVKEN
       ggrbp  ---------- -----MAYTW RALLLLALAF LGSSMAERDC RVSSFKVKEN
       xlrbp  ---------- -----MERKV LGL.LIALGF LGSCLAEKNC RVDNFEVMKD

                    4  5                                        6 7
              51
鱼类   ccrbp  FDRMRYQGTW YAVAKKDPVG LFLLDNVVAN FKVQEDGTMT ATATGRVIIL
       drrbp  FNRTRYQGTW YAVAKKDPVG LFLLDNIVAN FKVEEDGTMT ATAIGRVIIL
       omrbp  FDRSRYTGRW YAVAKKDPVG LFLLDNVVAQ FSVDESGKVT ATAHGRVIIL
       sarbp  FDKTRYAGTW YAVGKKDPEG LFLIDNIVAQ FTIHEDGAMT ATAKGRVIIL
其他   mmrbp  FDKARFSGLW YAIAKKDPEG LFLQDNIIAE FSVDEKGHMS ATAKGRVRLL
动物   rnrbp  FDKARFSGLW YAIAKKDPEG LFLQDNIIAE FSVDEKGHMS ATAKGRVRLL
       btrbp  FDKARFAGTW YAMAKKDPEG LFLQDNIVAE FSVDENGHMS ATAKGRVRLL
       ssrbp  FDKARFSGTW YAMAKKDPEG LFLQDNIVAE FSVDENGHMS ATAKGRVRLL
       ecrbp  FDKARFSGTW YAMAKKDPEG LFLQDNIVAE FSVDEYGQMS ATAKGRVRLL
       hsrbp  FDKARFSGTW YAMAKKDPEG LFLQDNIVAE FSVDETGQMS ATAKGRVRLL
       ocrbp  FDKARFAGTW YAMAKKDPEG LFLQDNIVAE FSVDENGHMS ATAKGRVRLL
       ggrbp  FDKNRYSGTW YAMAKKDPEG LFLQDNVVAQ FTVDENGQMS ATAKGRVRLF
       xlrbp  FNKERYAGVW YAVAKKDPEG LFLLDNIAAN FKIEDNGKTT ATAKGRVRIL

                  8          9                10      11
              101                                                       150
鱼类   ccrbp  NNWEMCANMF GTFEDTEEPA RFKMKYWGAA AYLQTGYDDH WIIDTDYDNY
       drrbp  NNWEMCANMF GTFEDTEDPA KFKMKYWGAA AYLQTGYDDH WIIDTDYDNY
       omrbp  NNWEMCANMF GTFEDTPDPA KFKMRYWGAA SYLQTGNDDH WVIDTDYDNY
       sarbp  NNWEMCADMM ATFETTPDPA KFRMRYWGAA SYLQTGNDDH WVIATDYDNY
其他   mmrbp  SNWEVCADMV GTFTDTEDPA KFKMKYWGVA SFLQRGNDDH WIIDTDYDTF
动物   rnrbp  SNWEVCADMV GTFTDTEDPA KFKMKYWGVA SFLQRGNDDH WIIDTDYDTF
       btrbp  NNWDVCADMV GTFTDTEDPA KFKMKYWGVA SFLQKGNDDH WIIDTDYETF
       ssrbp  NNWDVCADMV GTFTDTEDPA KFKMKYWGVA SFLQKGNDDH WIIDTDYDTY
       ecrbp  NNWDVCADMV GTFTDTEDPA KFKMKYWGVA SFLQKGNDDH WIIDTDYDTY
       hsrbp  NNWDVCADMV GTFTDTEDPA KFKMKYWGVA SFLQKGNDDH WIVDTDYDTY
       ocrbp  NNWDVCADMV GTFTDTEDPA KFKMKYWGVA SFLQRGNDDH WIIDTDYDTF
       ggrbp  NNWDVCADMI GSFTDTEDPA KFKMKYWGVA SFLQKGNDDH WVVDTDYDTY
       xlrbp  DKLELCANMV GTFIETNDPA KYRMKYHGAL AILERGLDDH WVVDTDYTTY
```

图 11-13 我们将会介绍一种通过对 13 个直系同源的 RBPs 进行多重序列比对来构建系统发生树的方法，该方法使用 PileUp 软件（基于 GCG）。只有靠近序列氨基末端部分的多重比对结果被展现出来。图 3-8 使用了与之完全相同的数据并且生成了在图 3-2 中所描述的那棵进化树（详细信息可参见原始图示）。其中有四个序列来自于鱼［用垂直的竖线来标注；ccrbp 是鲤鱼（*Cyprinus carpio*）的氨基酸序列；drrbp 是斑马鱼（*Danio rerio*）的氨基酸序列；omrbp 是虹鳟鱼（*Oncorhynchus mykiss*）的氨基酸序列；sarbp 是一种硬骨鱼（*Sparus aurata*）的氨基酸序列］。有 1 个 RBP 的氨基酸序列来自一两栖类的动物（*Xenopus laevis*）：*xlrbp*。其他 9 种氨基酸序列分别来自于小鼠的氨基酸序列：*mmrbp*。大鼠的氨基酸序列为 *rnrbp*，牛的氨基酸序列为 *btrbp*，猪的氨基酸序列为 *ssrbp*，马的氨基酸序列为 *ecrbp*，人的氨基酸序列为 *hsrbp*，兔的氨基酸序列为 *ocrbp*，以及小鸡的氨基酸序列为 *ggrbp*。在多重序列比对中所产生的缺口（参见图中的箭头 1）是很难简单地通过系统发生的算法来确定到底是由于插入还是删除产生的。有很多位点是 100%保守的（例如，箭头 2 所指示的半胱氨酸以及箭头 4 和箭头 5 所指示的 GXW 区域）。鱼序列中的很多氨基酸位点相对于其他很多物种的序列是很不相同的（参见图中箭头 3、箭头 6、箭头 8、箭头 11）或者在一组鱼和两栖类呈现差异性（参见图中箭头 7 和箭头 10）。一棵系统发生树为这些相互关系提供了一个可视化的图示（在图 3-2 和本章中将有更详细的介绍）

② 一些多重序列比对的程序（例如 PileUp 和 ClustalW），可能直接将相关序列排列在其他序列的外边。如果必须包含该序列，将会采用更低的缺口产生和/或缺口延长的惩罚机制来容纳多重比对中较远同源关系的序列。

③ 某些序列的完整序列是不可知的：对于虹鳟鱼（*Oncorhynchus mykiss*）和牛（*Bos taurus*）的 RBP 序列都没有起始的蛋氨酸位点。通常意义上来说，用于系统发生分析的多重序列比对数据应该严格限制在对于每一种所研究的分类都可获得的那段蛋白质（或者核苷酸）序列。

④ 在这些多重序列比对中存在着很多偶然的内部缺口（参见图 11-13，箭头 1）。一个缺口可能代表某些序列中的插入事件或者另外那些序列的删除事件。大部分系统发生算法是不预先估算删除和插入事件的（也统称为插入删除子：indels）。大部分专家倾向于认为应该将任何一组多重序列比对中任何位置上产生的缺口删除掉。

⑤ 在这个例子中，值得引起注意的是 4 种来源于不同鱼类的 RBP 序列（鲤鱼、斑马鱼、虹鳟鱼和一种雌雄同体的硬骨鱼）。直觉上，我们总是期望这 4 种鱼类序列比其他 9 种序列存在更多的差异性，这 9 种序列包括哺乳动物（例如人类）的氨基酸序列，一种两栖类动物（*Xenopus laevis*）的氨基酸序列，以及一种现已灭绝的鸡类（一种古蜥类，是包含鳄鱼和鸟类的那一界）的氨基酸序列等。因而，尽管所有的 13 种含有 RBP 的物种都是脊椎动物，我们还是期望鱼类的氨基酸序列应该和剩下物种的氨基酸序列存在较大的距离差异性，我们假设他们在亿万万年前来源于一个共同的祖先。确实，我们通过多重序列比对的结果可以来观察到这种差异性。在鱼和其他 9 种物种中的特殊位点上存在着一些不同的氨基酸位点。可参见图 11-13 中的例子（详见箭头 3、箭头 6、箭头 8 和箭头 11）。其他还有一些位置是在鱼和两栖类是相似的（参见图中箭头 7、箭头 10），其他还有很多位点在图中所有物种的蛋白中都是保守的。大部分位点在图中所有物种的蛋白质中均是保守的（可参见图中的箭头 2、箭头 4 和箭头 5），因此这些蛋白质被假定为一组具有较近亲缘关系的蛋白质家族。系统发生树（参见图 3-2）是这些亲缘关系的可视化表示。我们在任何时间从任何角度观察一多重序列的比对结果和一棵进化树，都能获得相似的信息。

⑥ 根据第 10 章的观点，生成一棵正确的系统发生树中最为重要的步骤就是产生一个好的多重序列比对结果。这在低保守性的蛋白序列之间是一件难度特别大的事情，因为其中有关那些含糊不清的缺口位置在不同的比对方法中得到的结果是截然不同的。大部分研究者认为，应该将那些含有含糊不清缺口位置的序列片段从多重比对中去掉。在这些序列区域中武断的比对决定往往对于系统发生分析的结果是有利的。

Joe Felsenstein（目前就职于华盛顿大学基因组科学系）提供了一个包含大约 200 个系统发生软件的 web 页面（http://evolution.genetics.washington.edu/phylip/software.html）。该网站包含了有关 PHYLIP 软件的很多相关信息（并提供一些有关 PHYLogeny 的补充软件包）（详见 http://evolution.genetics.washington.edu/phylip/general.html）。这是目前比较流行和常用的系统发生软件之一，并且它可以在任何主要的计算机操作系统上正常运行。

11.3.3 第三步：构建系统发生树的方法

构建系统发生树的方法有很多种，在很多书籍（参见 Nei 在 1987 年，Graur 和 Li 在 2000 年，Li 在 1997 年，Maddison 和 Maddison 在 2000 年，Durbin 等，在 2000 年，Baxevanis 和 Ouellette 在 2001 年，Clote 和 Backofen 在 2000 年，Hall 在 2001 年出版的学术著

Los Alamos 国家实验室的 HIV 序列数据库（我们将在本书第13章中展开详细讨论）提供了一个简要的有关如何建立和解释系统发生树的在线帮助文件（详见 http://hiv-web.lanl.gov/content/hiv-db/TREE_TUTORIAL/Tree-tutorial.html）。这个站点上有链接到 PAUP（将在下文中作详细介绍）、PHYLIP 和其他一些进化树的构建软件。

作）和研究文献（参见 Felsenstein 在 1996 年和 1988 年，Hein 在 1990 年，Nei 在 1996 年，Thornton 和 DeSalle 在 2000 年发表的研究文献）中具有详细的介绍。对于构建系统发生树需要注意的原则和方法主要有两种：基于距离的方法和基于字母的方法。

基于距离的方法是通过计算分子序列之间的距离来构建系统发生树的。现存一些距离计算矩阵，例如统计不同序列中氨基酸变化数目的矩阵或者距离打分矩阵（参见框 10-1）。基于距离的方法所耗的计算机时间较少，它们包含了多重序列比对中的操作分类单元（OTUs）的距离打分信息。这样，利用所有参加比对的蛋白质序列（或者核苷酸序列）的成对打分矩阵可以生成一棵进化树。其中主要的基于序列的方法是非权重的成对算术平均值法（UPGMA）和近邻法（NJ）。同源聚类组数据库（COG）（参见表 8-9 和第 14 章）是依靠基于距离的比对方法来确定基因相互关系的。

基于字母特征的方法也主要为两种：最大吝啬法（MP）和最大似然法（ML）。吝啬分析包括搜索所观察到的不同蛋白质（或者基因）在该进化树中经过最少数目的氨基酸（或者核苷酸）替换方式的途径。对于描述有阶级性相互关系的基因和蛋白质序列，吝啬法搜索方法比基于距离的方法要有用得多。系统发生分类学假设基因存在于一个相互嵌套的阶级关系中，这就映射出序列中共享的特征是呈等级状阶梯性分布在序列中的。用最大吝啬法构建出的进化树被认为是对分化自一共同祖先的蛋白质（或者基因）相互关系的最佳描述。

吝啬（parsimony，这个单词来源于拉丁文 parcere，表示宽恕，不伤害的意思）代表一个简单的逻辑公式假设。

我们可以通过观察图 11-13 中描述的多重序列比对结果来思考一下有关基于距离和基于字母特征的分子系统发生学的差别。在基于距离的方法中，我们可以计算出多重序列比对结果中每一对蛋白质序列的氨基酸相似百分比。一些序列对（如大鼠和小鼠的 RBP）的亲缘关系非常接近，并且在进化树上的位置非常靠近。其他序列（如 4 种鱼的 RBPs）与另外一些序列相比进化距离更远并且将会被放置在树上相隔较远的位置。直觉上，我们可以在水平角度考察图 11-13 中的序列信息，并在整个序列度量范围内计算进化距离。

相反地，基于字母的方法是从一个垂直的角度来分析多重序列比对结果的。在每一列氨基酸的排列形式中，哪一种排列方法是解释字符进化（例如氨基酸序列变化）最简单的（或者说是最简约）的方法？将鱼类放置在离其他蛋白质序列很远的进化树上的位置是通过对图 11-13 一些位点考察后的结果，参见箭头 3、箭头 6、箭头 8 和箭头 11。

ReadSeq 是 Don Gilbert 在 1993 年编写并让公众免费使用的。

无论是采用基于距离的还是基于字母特征的方法，大部分进化树构建软件都将多重序列比对结果作为其输入项。ReadSeq 是一个便利的在线多重序列比对结果翻译软件，它可以将多重序列比对的结果翻译成大部分常用的系统发生软件所识别的传统文件格式。一些 ReadSeq 的服务器地址罗列在表 11-3 中。

表 11-3 因特网上可用的 ReadSeq 服务器

来 源	超 链 接
Baylor 医学院	http://searchlauncher.bcm.tmc.edu/seq-util/readseq.html
国际健康研究所信息技术中心	http://bimas.dcrt.nih.gov/molbio/readseq/
巴斯德研究所	http://bioweb.pasteur.fr/seqanal/interfaces/readseq-simple.html
欧洲生物信息研究所	http://www.ebi.ac.uk/readseq/

注：由于有限篇幅还有很多没有列出。

11.3.3.1 用基于距离的方法来构建系统发生树

测定序列距离的方法有很多种，其中最简单的方法是比对成对的序列，并且统计出不同序列的数目。分化度也可被称为加权平均距离。对于序列长度为 N，差别位点数为 n 的比对，分化度 D 的定义公式如下：

$$D=\left(\frac{n}{N}\right)\times 100\% \tag{11-5}$$

加权平均距离计算起来非常简单，但是它却忽略了很大一部分序列进化关系中的信息量。其中主要的原因是字符差异并不一定和距离差异相同：两个序列的差异是很容易能被测量出来的，但是包含了很多不能直接被观察到的突变的序列间的遗传距离很难被简单地测量出来。例如图 11-11 中所示，通过统计序列的差异性来估算突变，可能会有很多实际上已经发生的突变不被察觉。Jukes 和 Cantor 在 1969 年提出一矫正公式：

$$D=-\frac{3}{4}\ln\left(1-\frac{4}{3}p\right) \tag{11-6}$$

Jukes-Cantor 的单参数模型描述了每个氨基酸突变为另一个氨基酸的概率（图 11-14）。这个模型有一个简单的假设，即每个残基改变为其他 3 种残基的概率相同，并且 4 种残基的出现频率也相同。因此，这个模型假定序列发生转换和颠换的概率相等。这个修正值对于亲缘关系非常接近的序列来说是非常小的，但是实质上对于那些亲缘关系更远的序列来说可能更接近于实际值。当然，当两个序列差异性超过 70%的时候，两者的距离差异性就很难被估算了。

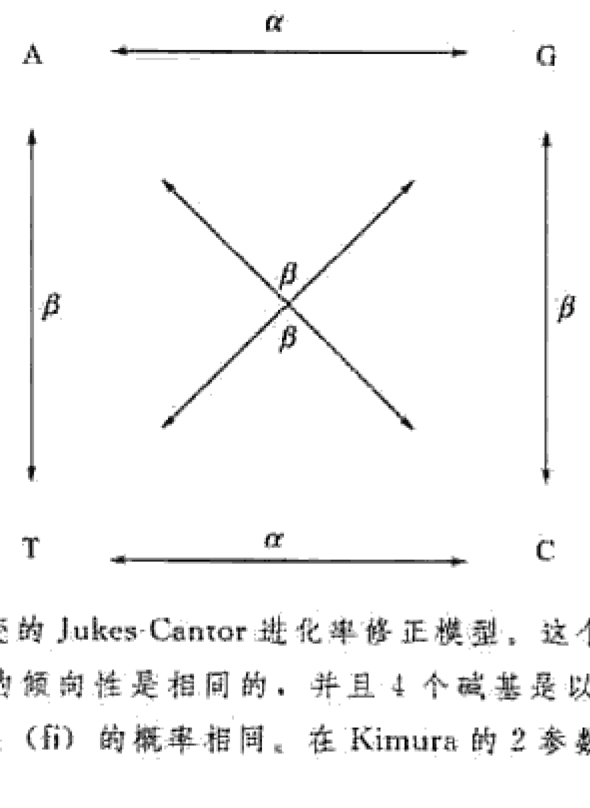


图 11-14 比对中阶层改变的 Jukes-Cantor 进化率修正模型。这个模型假设每一个核苷酸残基改变成其他氨基酸残基的倾向性是相同的，并且 4 个碱基是以相同的概率在序列中出现的。假设转换（ct）和颠换（fi）的概率相同。在 Kimura 的 2 参数模型中设定 $\alpha \neq \beta$。一般来说，颠换的所附权重较大

大多数情况下，转换的概率要比颠换的概率大得多。目前有很多模型比 Jukes-Cantor 的模型要成熟得多。Kirnura 的 2 因子模型通过赋予颠换更大的权重来调整转换和颠换的比值，以此来计算蛋白质编码区中引起非同义改变的似然值。在任何一个 DNA 的区域（包括非编码序列），转换与颠换的比值矫正了双螺旋结构中建立的嘌呤-嘌呤或者嘧啶-嘧啶的生物物理阈值。有一些矫正值是用来调整超出传统上限的变异率（例如，病毒或免疫球蛋白经常会出现一些超突变的变异区域）或者有些矫正值是用来矫正不同密码子位置的突变率的。

基于进化距离构建系统发生树的方法有一个与生俱来的缺陷，即其抛弃了字母特征的数据。一个典型的例子就是，如果两个比对序列是在一个特殊的残基上存在差异性，那么基于距离的方法一般情况下将无法解释以后的突变是否还是会倾向于发生在这个位点上。同样，基于字母特征构建系统发生树的方法（如吝啬算法）也抛弃了有关距离信息的数据并且做了相应的简化假设（在下文中将有详细介绍）。下面举例说明如何实际应用方程 11-6。假设有一个比对结果是 60 个比对残基中存在 3 个差异位点，那么其归一化的加权平均距离为 3/60＝0.05，相应的 Jukes-Cantor 模型矫正值为：$d=-(3/4)\ln[1-(4\times 0.05/3)]=0.052$。在这个例子中，应用这个矫正值只是起到了很小的效果。但是当序列中存在 30 个氨基酸的差异性时，Jukes-Cantor 矫正值变为一个真正起较大矫正作用的值：$-(3/4)\ln[1-(4\times 0.5/3)]=0.82$。

11.3.3.2 利用基于距离的 UPGMA 方法来构建系统发生树

PAUP 程序是 David Swofford 独立完成的，他目前就职于佛罗里达州立大学。该软件可以在 Sinauer 协会网站上（http://www.sinauer.com）得到，并且可以在大多数计算机软件平台上使用。该程序自带非常有用的使用指南。更多的信息可参见 http://paup.csit.fsu.edu/。在本章中我们将主要介绍 Macintosh 操作系统下运行的 PAUP 程序。

目前有很多流行软件程序可以被用来构建系统发生树。在下文中我们首先来关注用吝啬算法来建立的系统发生分析的程序：PAUP。这是目前最通用、最全面和最流行的可以用来构建系统发生树的软件之一。数据可以以各种多序列组合（multiple sequence alignment）格式导入，分析一般可能采用简约性分析、距离分析或者其他的标准（图 11-15）。我们可以在 PAUP 程序中简单地通过在分析菜单中选择距离标准后，再选择 UPGMA 来构建一颗基于距离特征的进化树［图 11-15(d)］。

此时会弹出一个对话框，让你选择是使用 UPGMA 还是近邻连接（NJ）算法［图 11-16(a)］。UPGMA 算法基于一个距离矩阵的序列聚类，是一种简单的构建系统发生树的算法。我们首先着重介绍 UPGMA 的原因是其构建系统发生树的过程较为直观。但是，最专业的构建基于进化距离的系统发生树的方法还是近邻连接（NJ）算法（将在下文中给予介绍）。随着序列聚类的增加，一棵进化树被组装起来。图 11-16(b) 描述了用 UPGMA 算法构建的由 13 个 RBPs 组成的系统发生树。正如我们预期的一样，4 种鱼类的 RBP 序列是与其他物种分离开来的，形成一个清晰独立的进化分支。两个进化距离最为接近的关联蛋白质，大鼠和小鼠的 RBP 共享了 99.5%的氨基酸同一性。它们相互之间聚类得非常接近。值得引起注意的是哺乳动物的 RBPs 也构成了一个独立的进化分支，与鸟类和两栖类的蛋白质分列在不同的进化分支上。

UPGMA 算法相对较为简单。下面我们来考虑进化距离能够用图线上的几个点来代表的 5 种序列［图 11-17(a)］。某些蛋白质序列（例如序列 1 和序列 2）是有高度相似性的，

另外的蛋白质序列（例如序列 1 和序列 3）相互之间的相关性很小。UPGMA 算法将这些序列聚类成如下的形式（数据来源于 Durbin 等 1998 年的著作）。

> 我们已经在第 10 章中详细介绍了如何利用一个距离矩阵来建立一棵引导树。

① 首先，每一个序列 i 都被分配给它们自己的聚类 C_i。在我们的例子中，共有 5 个独立的聚类。每个都被放置在一个树末端的操作分类单元 OUT 上（或者说是一个叶节点上）。例如，我们将会在 x 轴上标注 1～5 来代表一棵树的最终节点。

② 两个聚类 C_i 和 C_j 的分类是通过它们两类之间的最短距离来定义的，定为 d_{ij}。如果两类之间存在很多距离相等的最短对，那么将随机选择其中的一对来计数。在我们的例子中，蛋白质

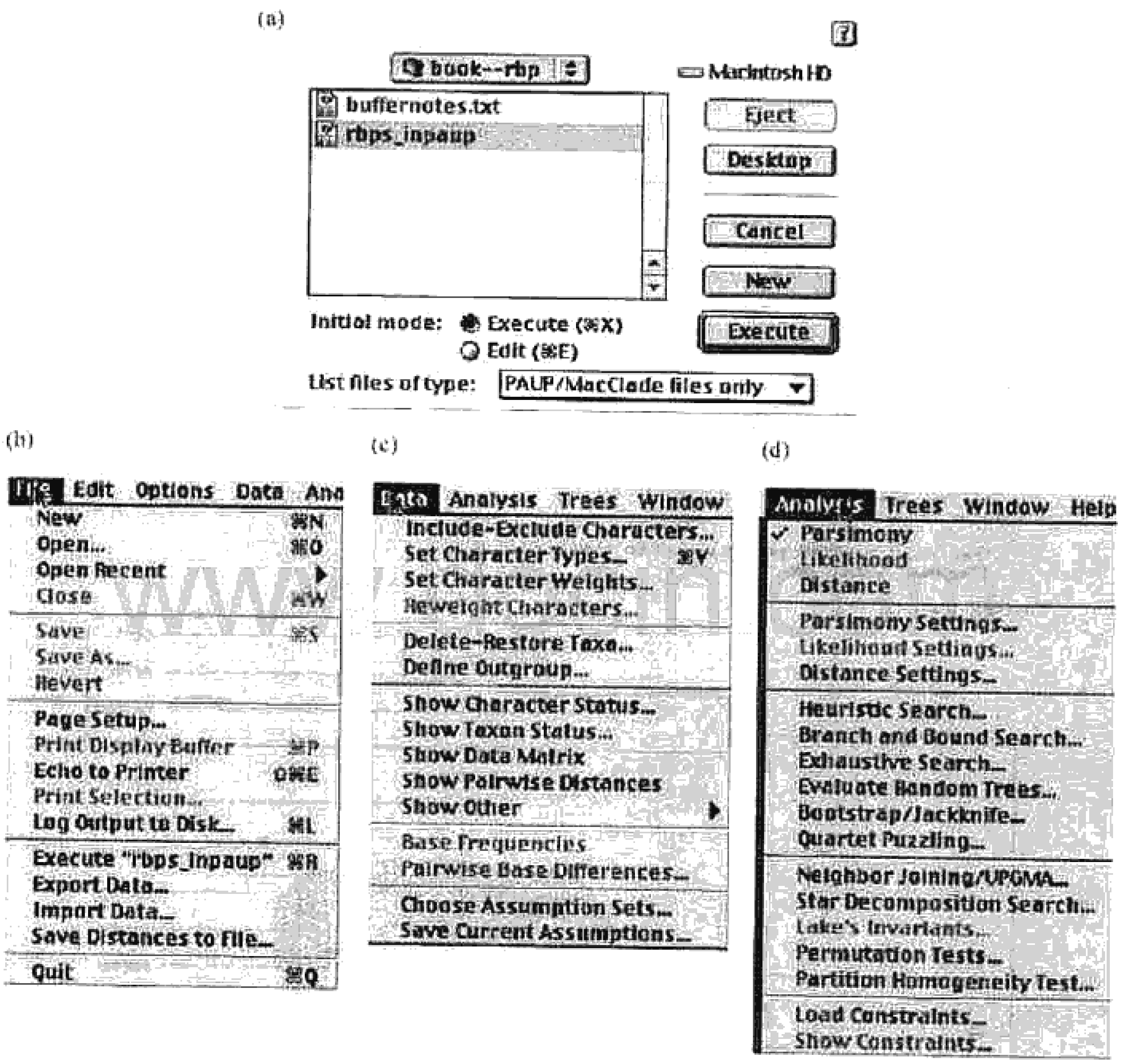


图 11-15 Smithsonian 研究所 David Swofford 撰写的 PAUP 程序，该程序提供了一个全面的研究系统发生的工具。可以用遗传计算学会（GCG）的多重序列比对的结果形式（MSF）或者 ClustalW（详见第 10 章）的结果来作为输入项，图中（a）为一对话框。（b）为主要的文件菜单。（c）一旦数据被输入和执行后，该数据可以用多种方式来进行分析。例如，通过“包括-拒绝字母特征”选项，我们可以很容易地对被选择的蛋白质（或者核苷酸）序列比对区域做系统发生分析。（d）其所带的分析工具允许使用者选择吝啬算法和基于距离的系统发生分析中任何一种方法

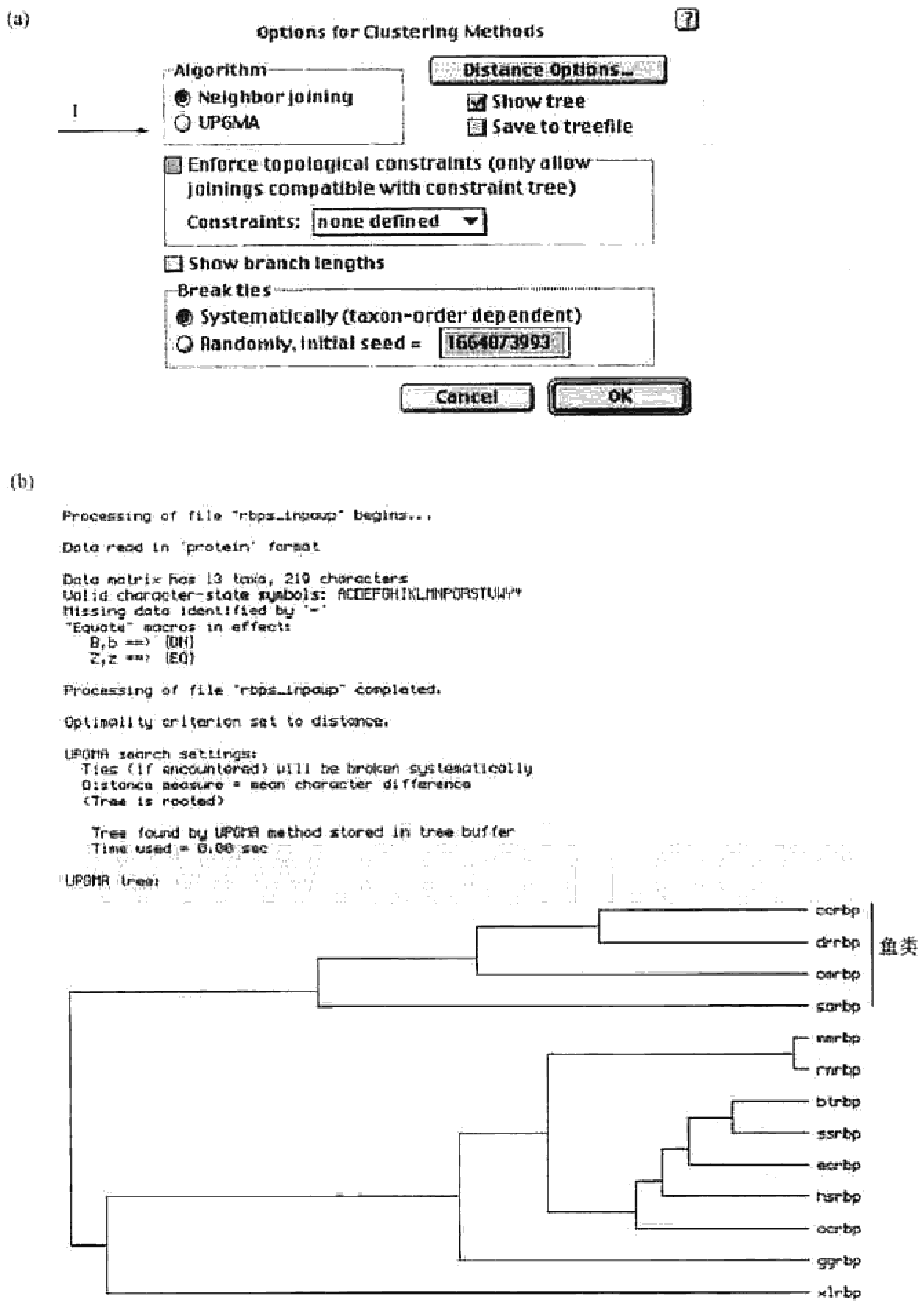


图 11-16 基于距离的最简单的构建系统发生树的方法之一是 UPGMA。(a) PAUP 中的工具通过将分析菜单置于"距离"选项 [详见图 11-15(d)]，然后选择 UPGMA 的选项。(b) 产生的进化树代表了 13 种 RBP 蛋白。值得注意的是，4 种鱼类的 RBP 蛋白和我们所期望的一样都是聚集在一起的。小鼠和大鼠的序列在进化树上的位置是相对较为接近的；这两个蛋白质共享了 99.5%的序列相似性

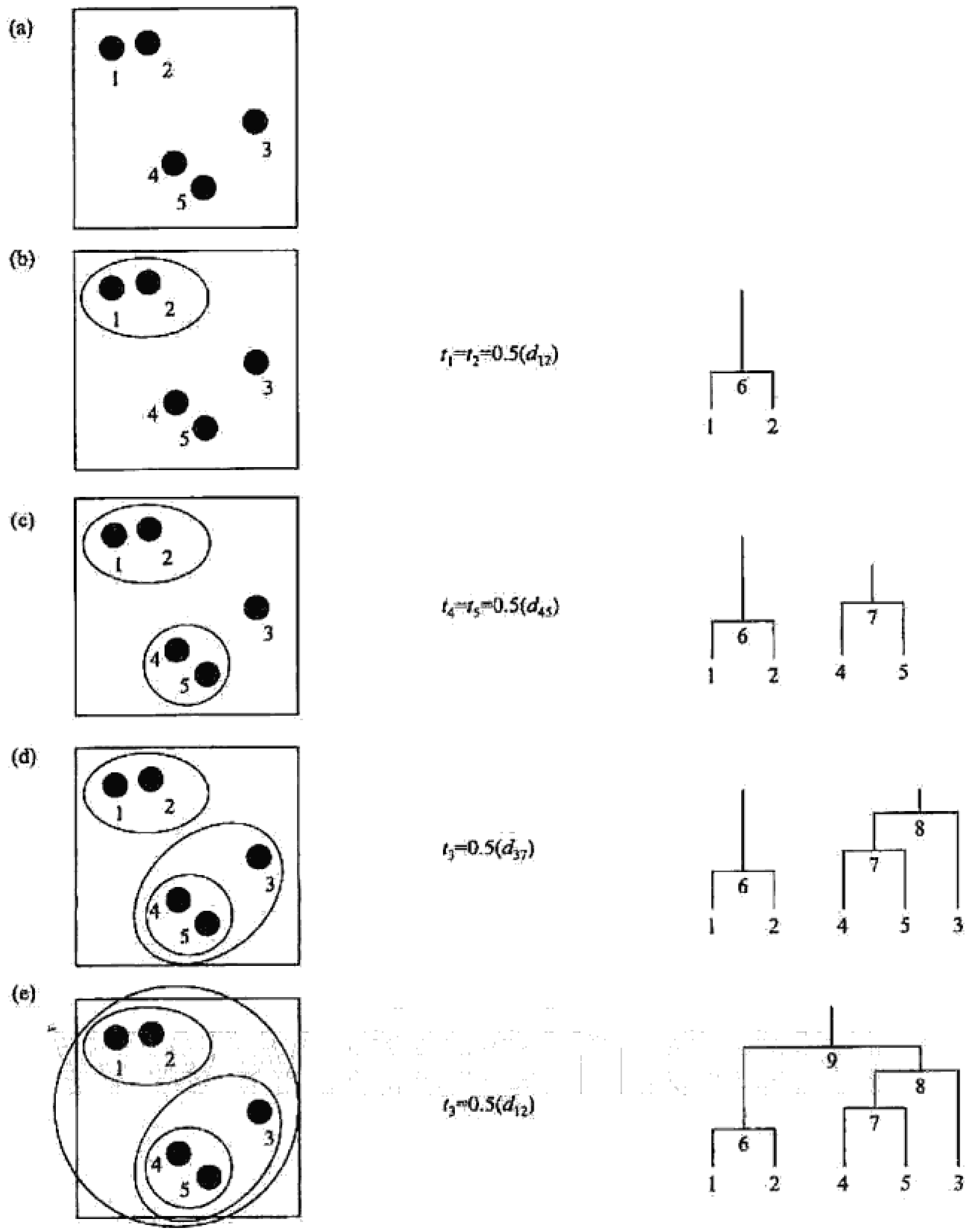


图 11-17 UPGMA 算法的图示。这是一个简单而且快速构建系统发生树的算法。它是一种基于聚类序列的算法。(a) 每一个序列都被分配给它自己的聚类。(b) 有最近距离的分类（序列 1 和序列 2）被确定和连接起来。这使得我们可以命名一个内部节点［参见图右边的节点 6，图 (b)］。(c) 下一个比较接近的序列是序列 4 和序列 5，然后 (d) 新形成一个组（序列 4 和序列 5）和序列 3。最后，(e) 中所有的序列都在进化树中的相应位置得到连接

序列 1 和序列 2 是第一对被聚类起来的序列［图 11-17 (b)］。值得注意的是我们可以马上定义一个内部节点。我们将它标记为 6，因为 1～5 是预留给进化树外部的叶节点的。

③ 一个新的聚类 k 被定义为 C_k，其中 $C_k = C_i \cup C_j$。

一个新节点 k 是通过女儿节点 i 和 j 来定义的，并且它被放置在进化树 $1/2d_{ij}$ 的高度处。对于新产生树的步骤中 i 树节点和 j 树节点被去除并用聚类 k 来代替。

④ 就这样循环往复前进，直到每一个新的聚类都被重新定义，并且计算出一个新的距

离矩阵 C_k 并替代原来的行 C_i 和 C_j，然后通过计算 C_i 和 C_j 的平均距离值来计算新的距离。在我们的例子中，$d_{3(1+2)}=(d_{31}+d_{32})/2$。

⑤ 当只剩下两类（i 和 j）保留下来时，树根节点被放置在树 $1/2d_{ij}$ 的高度位置。

用 UPGMA 建立起来的系统发生树通常是有根的。UPGMA 方法的一个边界条件是对于树中所有序列其分子时钟都是恒定的。倘若这个假设正确，那么进化树分支的长度可以被用来估算物种分化的时间。如果这个假设不正确，并且树的不同分支的替换速率是不相等的，那么这个方法将会产生一棵不正确的进化树。

在这里大篇幅介绍 UPGMA 方法是因为它类似于一个基于距离的方法。而且，它也被大量的应用在芯片数据的分析中（框 11-3）。David Hillis 和其他一些研究者研究证明：对于分子序列数据用这个方法比用连结近邻的方法准确率低很多。

框 11-3 去除芯片数据与构建系统发生树

> 用芯片数据构建的系统发生树（例如图 7-10 和图 7-12 中所示的进化树），每个元素之间是非常相似的，并且都连接在进化树一些较短的分支上。长的进化分支是与非相似性相对应的。系统发生树是用 UPGMA 或相似的算法来构建的，同样地，更远的距离代表对象之间有较长的分支。总体上看，可能这些树相互之间看起来都非常相似。一个较为可靠的基于距离的方法是将几乎所有的矩阵都用来构建系统发生树。所以，这些算法被应用到各种形式的数据中。

11.3.3.3 用基于距离的方法来构建系统发生树：连接近邻法（NJ）

> 任何一对操作分类单元都可以形成第一对聚类，因此一共有 $0.5N(N-1)$ 对可能的邻居。

NJ 的方法是用基于距离的方法来构建系统发生树（Saitou 和 Nei，1987）。我们利用星形的结构来开始生成一棵系统发生树［图 11-18(a)］。利用成对的比较来确定两个亲缘关系最为接近的序列。存在着一些邻居，彼此之间被定义为操作分类单元，并且通过一个单独的节点来连接［图 11-18(b)］。不断循环这个过程，直到遍历完整棵树的拓扑结构：相邻的节点 1 和节点 2 形成一个进化分支，并且在下一步中与操作分类单元 3 相互配对。

NJ 算法将每一步聚类得到的操作分类单元的分支长度的总和最小化。因此，有些时候该方法又被称为“最小进化”方法。它产生了树的拓扑结构和对树的分支总长度的估算。用 13 个 RBP 蛋白序列构建的 NJ 树在图 11-19 中有详细的介绍。当研究大数目分类的时候，这个算法是非常有用的。

> Los Alamos 国家实验室的 HIV 序列的数据库提供了一个在线的系统发生树的构建软件：Neighbor TreeMaker。它将多重序列比对的结果当作输入项，然后利用 PHYLIP 程序的 Dnadist 距离矩阵程序和 Neighbor 树文件生成软件来处理数据，最后打印出所产生的系统发生树。详细内容请访问如下网址：http://hiv-web.lanl.gov/content/hiv-db/CONTAM/TreeMaker/TreeMaker.html。

11.3.3.4 基于字母特征构建进化树的方法

基于字母特征的方法是通过对利用序列比对得到的亲缘关系所构建的候选树的进一步分析来构建进化树的。这些方法与基于距离的构建方法存在很大的不同，因为前者并不包括一个距离矩阵的中间概要形势。但是，它们还是基于距离规律的。

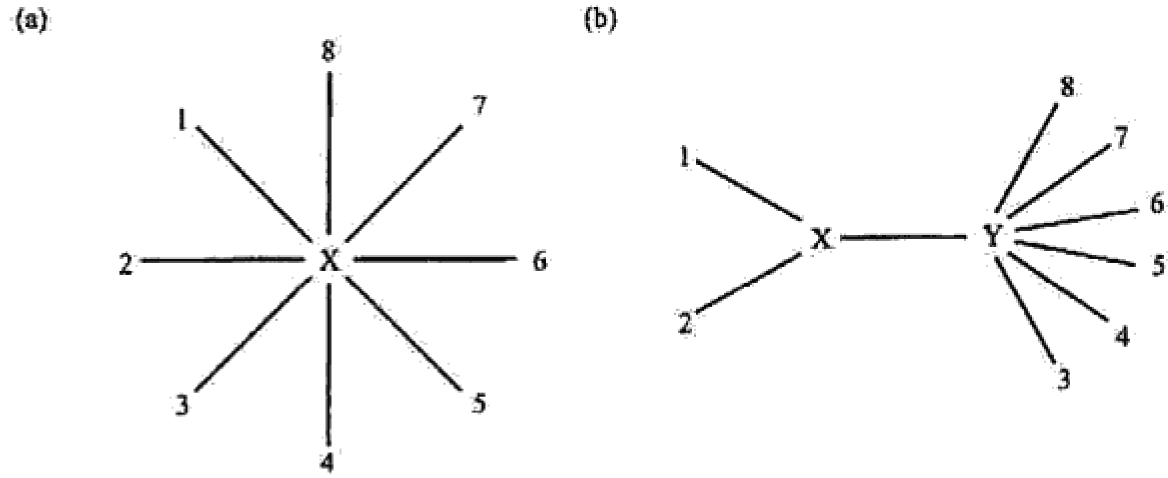


图 11-18 NJ 方法是一种基于距离的建树算法。(a) 所有的操作分类单元首先聚类成一个星形树状结构。“邻居节点”被定义为由无根的分叉树内部的节点来独立连接的操作分类单元 (OTU)。(b) 两个距离最近的操作分类单元 (OTU) 被确定，例如操作分类单元 (OTU) 1 和操作分类单元 2。这些邻居通过一条内部的边 XY 与其他的操作分类单元 (OTU) 相连接。(b) 中的操作分类单元 (OTU) 被选作邻居的那些是被认为能产生分支长度总和最短的聚类。这个过程循环往复直到整棵进化树产生。数据来源于 Saitou 和 Nei 1987 年的研究成果，并授权使用

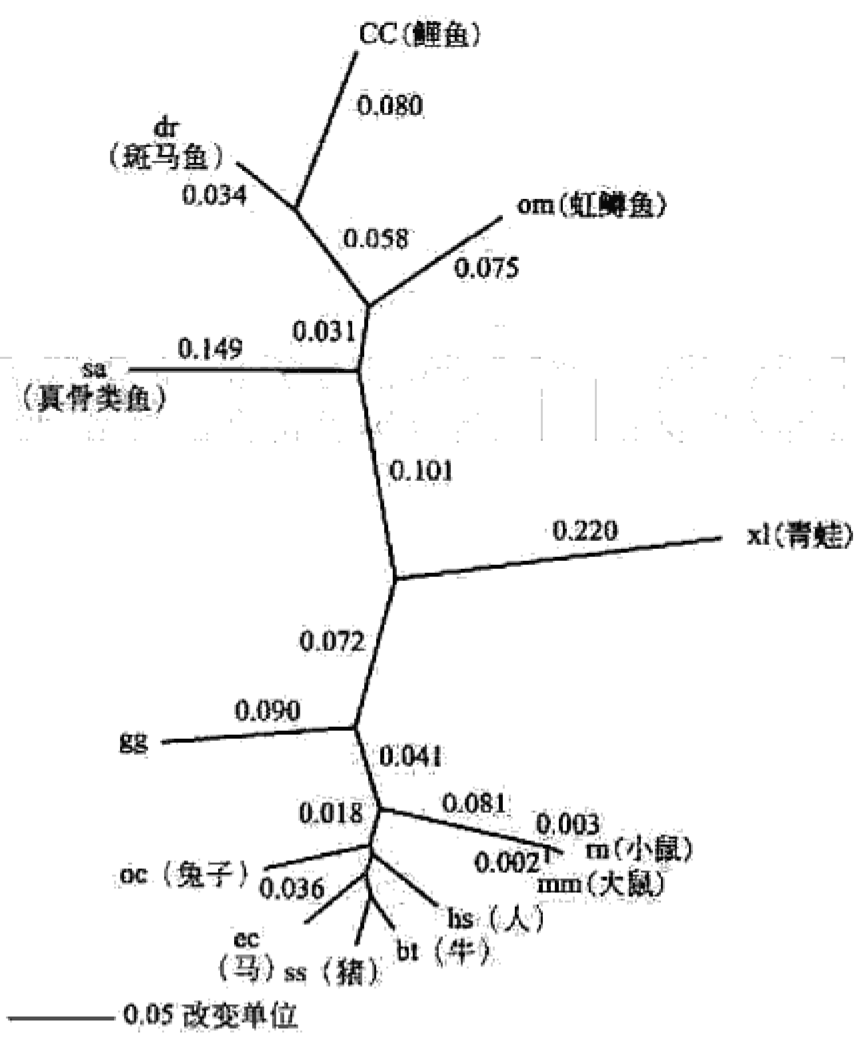


图 11-19 用距离标准的 NJ 方法来对 13 种 RBPs 做系统发生分析。这个分析工作借助 PAUP 程序来完成，并产生一棵无根树的系统发生图。为了更简洁地表示该进化树的关系，图中省去部分距离值的标注。缩写和新添的数值详情可参见图 3-2

11.3.3.5 最大吝啬法（MP）

MP 算法的主要思想是：拥有最短可能分支长度的进化树将是最好的系统发生树（Czelusniak 等，1990）。根据 MP 理论，用最少的改变步骤的方法来解释分子序列的进化方式比其他更为复杂的方法都更为有效。这里我们将为所观察得到的数据寻找一种最简约的解释。其中各步骤如下所示。

① 确定信息位点。如果一个序列位点恒定不变（例如图 11-13 中的箭头 2、箭头 4 和箭头 5），那么它就不是信息位点（详见下文叙述）。比对中的位点 43（右边箭头 3 所指的两列序列）也是非简约的信息位点，尽管它们的字母特征不是恒定不变的。其中 11 个分类含有丝氨酸，一个含有丙氨酸、另一个含有天门冬氨酸。所谓简约信息的位点，其不同字符特征必须各自至少在两个分类中单独出现。

② 每构建一棵系统发生树都会产生一个成本价值，那棵具有最低成本的进化树往往会被采用。当估算得到一个较为合理的分类树时（大概为一打或者更少），所有可能的树都将被估算并且选择其中具有最短分支长度的那棵。如果需要，可以借助启发式算法忽略系统发生树上那些不太可能包含最短分支树的大蛋白家族以降低搜索的复杂度。

③ 统计其中改变的数目并且选择其中最短的系统发生树（或者树群）。

为了说明 MP 算法是如何工作的，让我们以 4 个已经排列成行的氨基酸序列（图 11-20）为例。有 3 种可能的树可以用来描述这些序列之间的关系；所有的树都对应一个事先假定供比对的祖先序列。其中的一棵树（见图 11-20，左下角）只需要最少的步骤来解释观察序列是如何从一个假定的祖先序列进化而来的。在这个例子中，每个位点都是被单独加工的。带权重的和不带权重的简约算法都被发展了；权重是用来给特殊的核苷酸或者氨基酸替换事件增加代价的。例如，核苷酸中发生颠换的序列比发生转换的序列得到更大的罚分。

在 PAUP 程序中，你可以设定构建进化树的简约准则［图 11-15(d)］。运行穷举算法来遍历所有的进化树以此来找到分支总和长度最短的进化树的方法可能只能是一个美丽

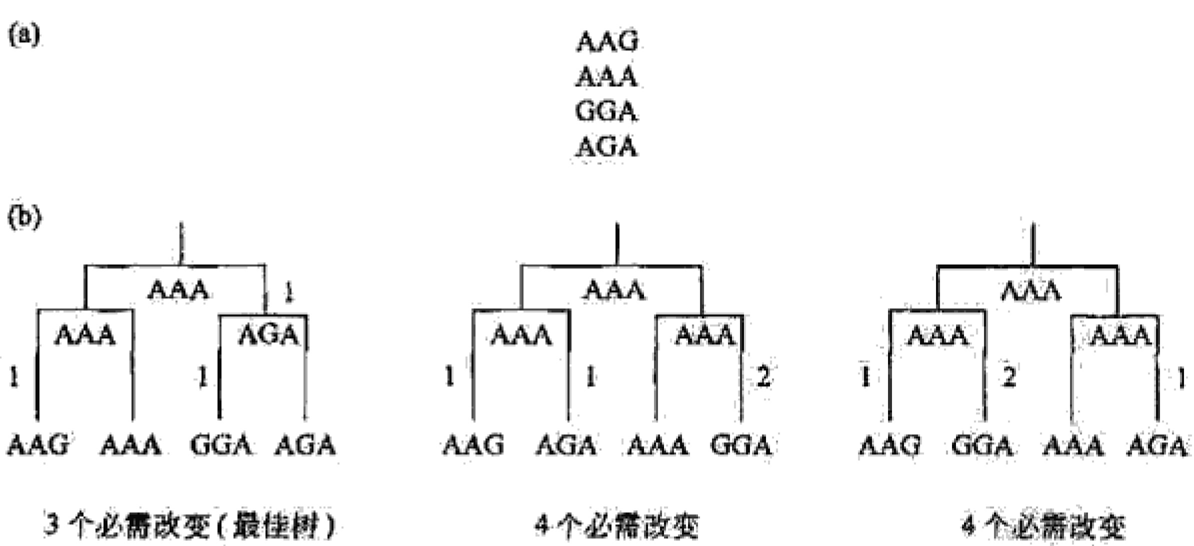


图 11-20 MP 算法的原理：(a) 考察 4 个比对的分类（例如序列信息），每个分类都是 3 个核苷酸的长度。3 棵不同的进化树模式描述了这些序列可能的进化方式。(b) 那棵需要最少改变步骤的进化树（在最左边的那棵树）将被 MP 算法所选择。数据来源于 Durbin 等 1998 年发表的研究论著，授权使用

的幻想。事实上，当分类数大于 10 或 12 时这种做法几乎是不可能的，因此就非常有必要引入一个启发式搜索的方法（图 11-21）。无论是启发式搜索算法还是穷举搜索算法最后往往会停留在对一些具有相同最短分支总和长度值的进化树鉴定和分辨上。这些树的群落可以被看作是系统发生图［图 11-22(a) 和图 11-22(b)］或者进化分支图。在系统发生图中，分支的长度是成比例的。因此，那些在系统发生树上相对位置较为接近的分类拥有较强的进化相关性。在进化分支图上情况相反。这样的图形安排有利于看清楚所有分支的权值，以及整洁的排列上有关节点的标签［图 11-22(d)］，尽管这样将减弱视觉效果。

一个人造的被称为长程吸引（long-branch attraction）的现象在简约算法中有时会发生。在一个蛋白质或 DNA 序列的系统发生重构中，一条分支的长度代表了两个分类之间发生替换的氨基酸数目。简约算法假定所有的分类都是以一个相同的进化速率来进化的，并且所有的字母特征的改变对总信息量改变的贡献也是相同的。长程吸引（long-branch attraction）指在进化树上将一些分类放置在一起是因为它们的进化速度较快而不是因为它们互相之间的亲缘关系较近的一种现象。这样聚在一起的分类往往含有人为的倾向，因为它们的序列上都含有很多突变。想像一下在图 11-23 中所描述的真实的进化树中所存在的现象，其中分类 2 代表一个相对于分类 1 和分类 3 而言进化更为迅速的蛋白质或 DNA 序列。树的外围部分

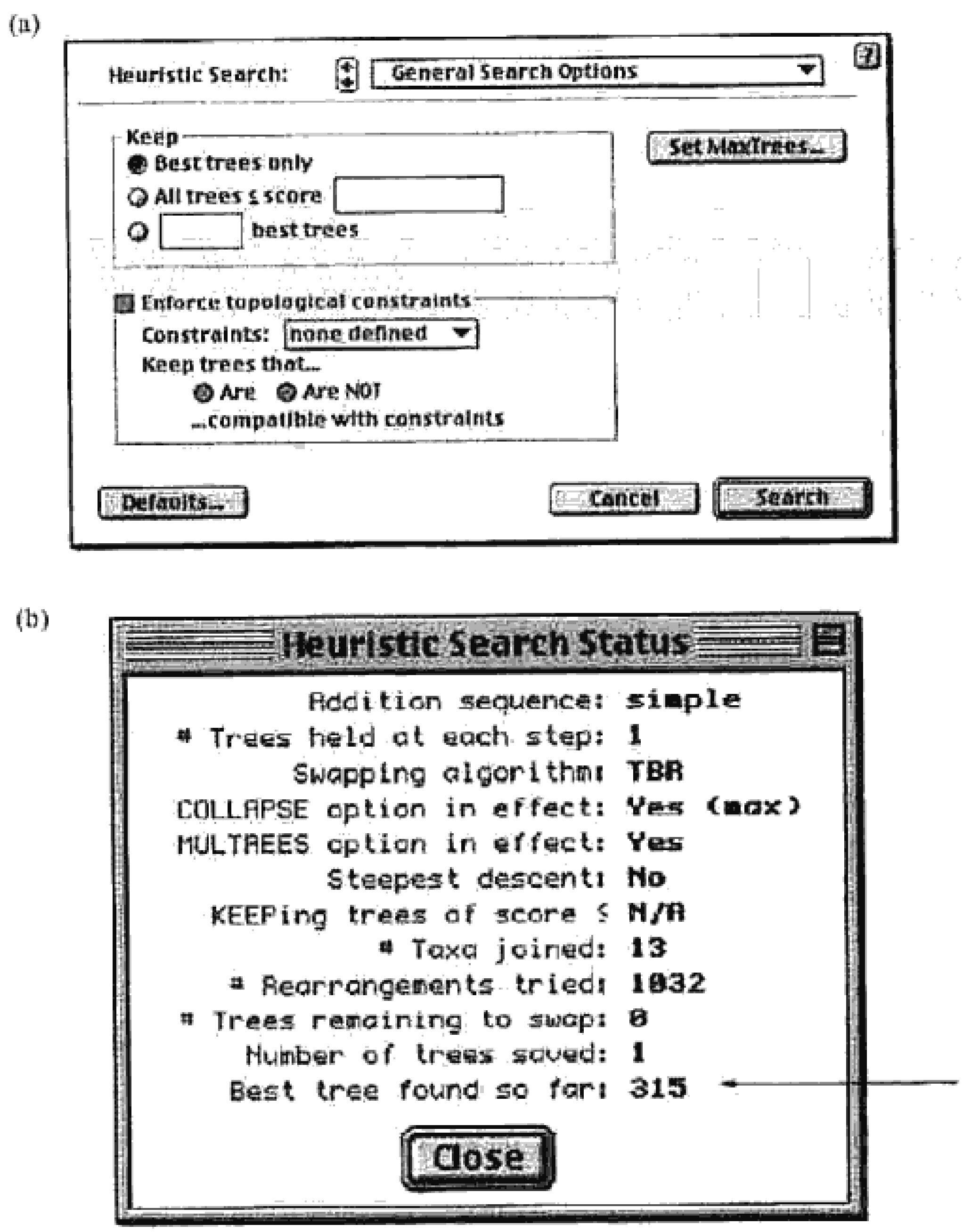


图 11-21 (a) PAUP 程序也可以运行启发式搜索。(b) 最佳树的总分支长度被报道（参见箭头所指部位）。用简约准则，可能可以发现多个进化树共享一个相同的最低分的现象。这些最佳进化树中被最终选定的那棵树通常都包含多个分支节点

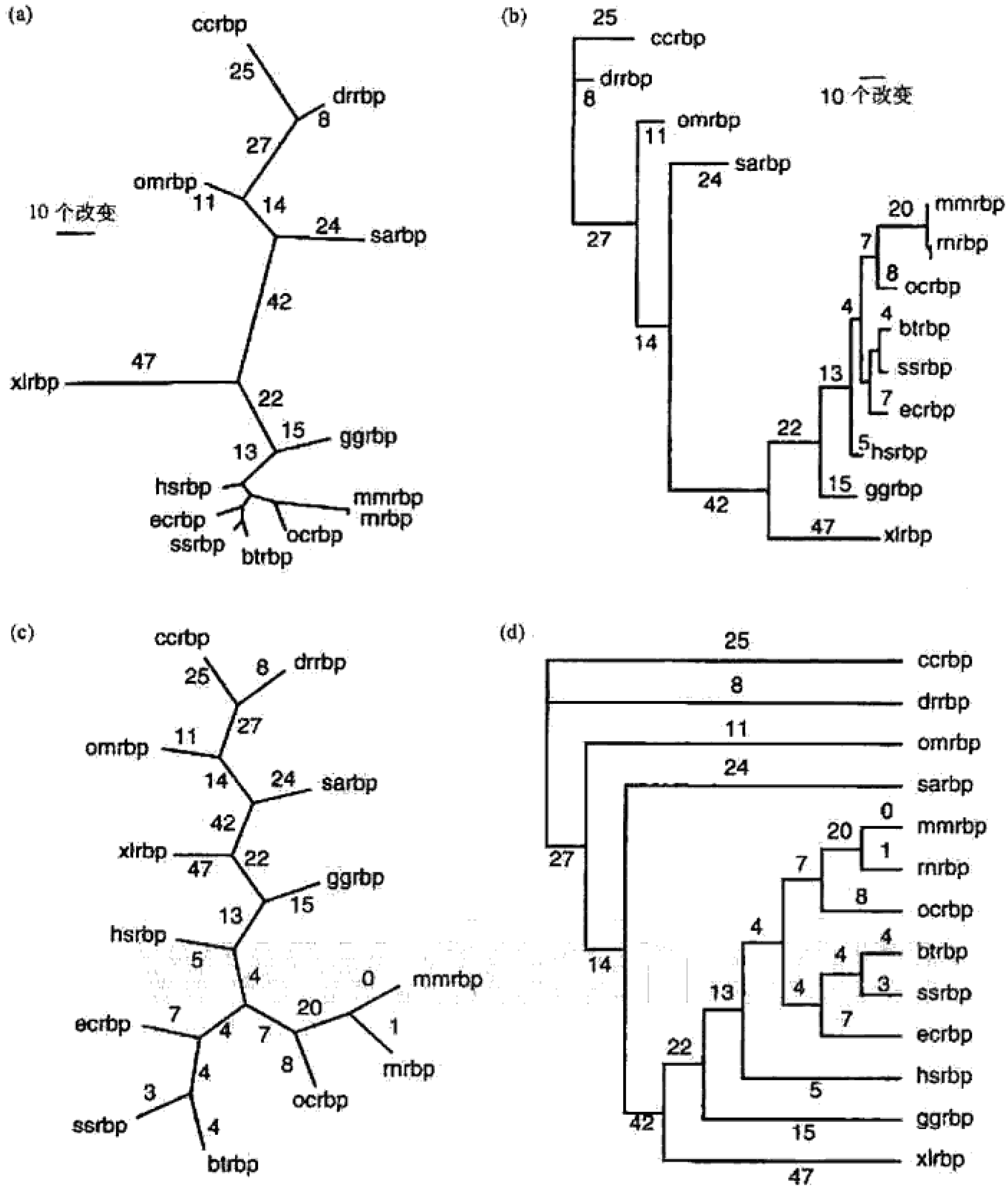


图 11-22 用简约准则来对 13 个 RBP 蛋白作系统发生分析。这些相同的数据分别被表示为如下的一些形式：(a) 系统发生图，(b) 矩形的系统发生图，(c) 进化分子图，以及 (d) 矩形的进化分支图。每种树的描述形式都有各自不同的优点。(a) 和 (b) 中的进化树其权重值与进化分支的长度是成比例的（并带有标度）；但是考虑到视觉问题这两棵树上的分支权重值不是全部都被显现出来。(c) 和 (d) 中的进化树其权重值与进化分支的长度不成比例；其中的数值比 (a) 和 (b) 中表示得更能够让人理解，有很近亲缘关系的序列被聚类在一起

（通过距离精确度来定义）比分类 1、分类 2 和分类 3 相互之间的亲缘关系要远得多。MP 算法可能会生成一棵推断树（图 11-23)，在这棵进化树上分类 2 被“吸引”到另一条较长的分支位置上（位于外围部分），因为这两个分类都有较高的替换率。任何时间只要有两个较长的进化分支出现，他们之间是可能会被互相“吸引”的。



图 11-23 长进化分支链的吸引现象。真实的进化树包含一个比其他分类进化速率都要快很多的分类（图中标注为 2）。其中分类 2 和分类 3 共享了一共同的祖先，但是在我们推论得到的进化树上，分类 2 被放置在与其他分类分离开的位置上，因为它被一个含有长进化分支的外围集团所吸引。数据来源于 philippe 和 Laurent（1998）

11.3.3.6 最大似然法（ML）

最大简约方法可能某些时候会得到较为失败的结果，特别是当进化树不同分支间的进化改变数目较大时。一个可选的基于字母特征的方法是最大似然法（ML）（由 Felsenstein 于 1981 年提出）。最大似然法（ML）是最耗计算机时的方法之一，但是也是最具灵活性的可选方法之一。最大似然算法（ML）的目的在于寻找一棵对于产生的观察值有最高可能性的进化树。似然值是依据一些替换过程的模型来对比对中的每个碱基都进行估算后得到的。在这个简单的例子中，这个模型认为所有的核苷酸和氨基酸改变的可能性都是相等的。最大似然法（ML）通常都包含一个统计不同替换率的参数。例如，TREE-PUZZLE 程序（Strimmer 和 von Haeselser，1996；Schmidt 等；2002）就是用来计算异质杂交点改变率的。

Korbinian Strimmer 和其研究组编写的 TREE-PUZZLE 程序可在如下网址得到：http://www. tree-puzzle. de/。也可以利用 PHYLIP 中的 DNAML 程序和 PAUP 程序来运行最大似然算法（ML）。Posada 和 Crandall 在 1998 年编写的 MODELTEST 程序利用分析 56 个 DNA 进化模型的记录似然值来选出其中最合适的进化数据。详情参见 http://inbio. byu. edu/Faculty/kac/crandall-lab/modeltest. htm。

11.3.4 第四步：用随机测试和引导检测的方法评估进化树

当构建出一棵系统发生树后，你是如何评估其准确度的？评价正确率的主要准则是密度、效率和鲁棒性（Hillis，1995；Hillis 和 Huelsenbeck，1992）。有些研究者着重于研究构建进化树方法的精确性或一棵特殊的系统发生树的精确性。其中最为常见的方法是引导检测的分析方法（Hillis 和 Bull，1993）。引导检测的方法并不是一个估算进化树精确性的技术手段。作为替代，他描述了进化树拓扑结构的鲁棒性：在给定一种特殊分支顺序的前提下，该构建进化树的算法是如何一步一步利用随机排列的原始数据集找到这种分支顺序的？

精确度代表了推论所得的进化树与真实进化树之间的相似性。

非参数估计的引导检测法如下所示。利用某些构树方法，将多重序列比对的结果作为产生系统发生树的输入数据。之后，程序自动构建一个与原始数据集大小完全相等的人工数据集，这个数据集是通过随机抽取多重序列比对结果中的分类来构造的。这通常是通过替代步骤来实现，也就意味着独立的分类可能在数据集中出现多次（或者根本就未曾出现过）。然后，用随机数据集产生一棵进化

带参数估计的引导检测方法是表示在不替换原始取样的前提下对数据集重复随机取样的检测方法。其结果是一些较短序列的比对对原始数据的依赖性更小，这种方法的使用频率远小于非参数引导检测方法。

树。这样产生了大量的靠引导检测方法构建的进化树，大概有100～1000多棵新的进化树是用这种方式来构建的。然后将这些引导检测产生的进化树与原始的、推论出来的进化树作一比较。我们从引导检测方法中得到的信息就是实际观察到的原始系统进化树中每个进化分支所出现的频率。

在PAUP程序中使用引导检测方法的过程参见图11-24中的相关信息。原始系统发生树进化分支上所给定的百分频率值是根据引导检测对其拓扑结构的支持率来确定的［图11-24(b)］。一般情况下，当引导检测值大于70%时该进化分支的存在被认为是被引导检测所支持的。该值是Hillis和Bull于1993年将统计显著性控制在$P<0.05$的水平下所推导出来的。该方法检测了随机权重特征对原始数据矩阵所起的作用，提供了一种如何测量由多重序列比对结果所产生的进化树上系统发生信号强弱性及其分布的方法。

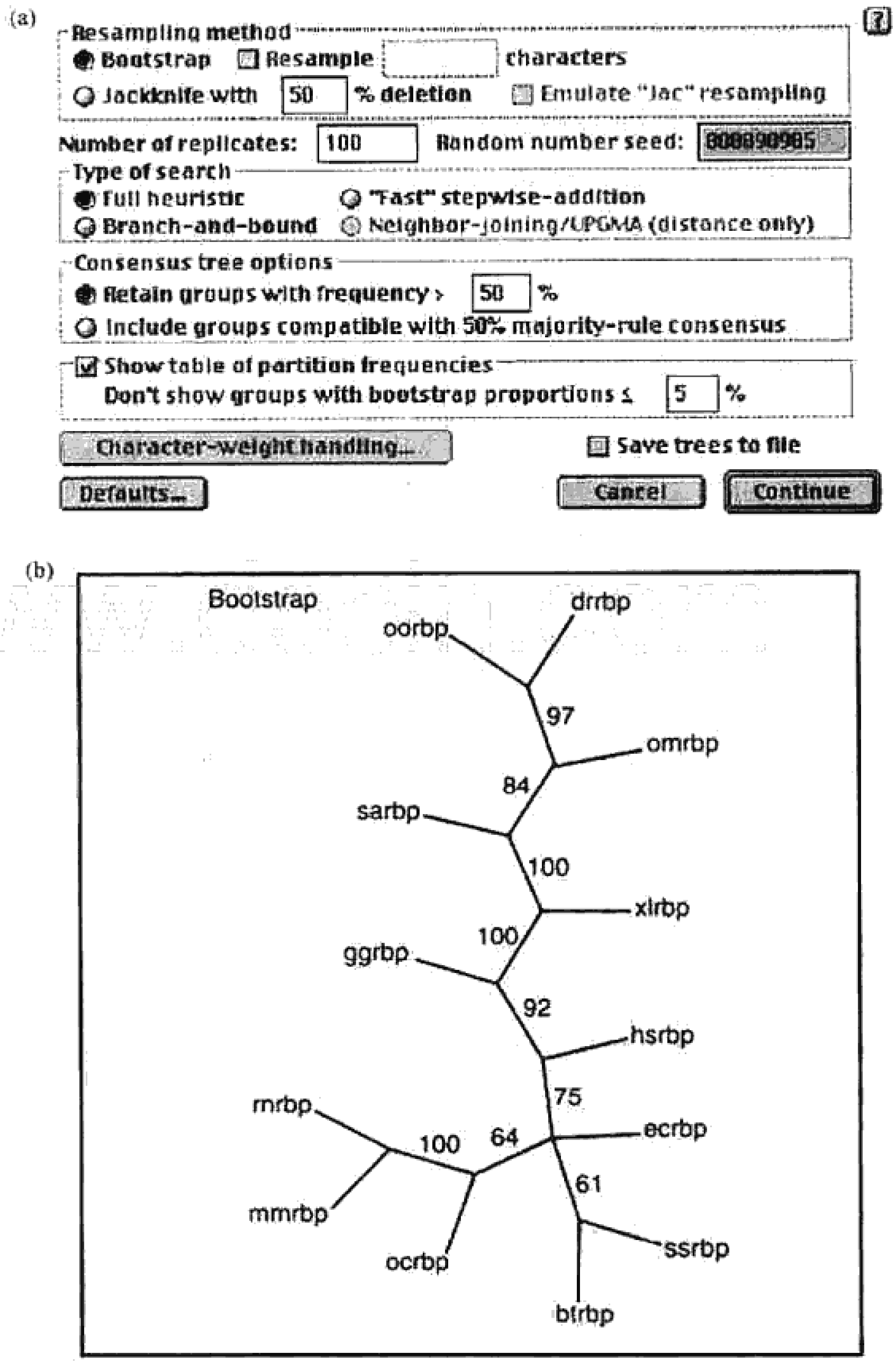


图 11-24 (a) PAUP程序擅长于自举方法。Jackknife是一种经常产生相似结果的补充重取样方法。通常我们选择运行100～1000个引导复制。(b) 作为结果输出的进化树，其每条进化分支上的值都表示引导树对其推论树分支支持的百分比。例如，在100%的引导测试中，小鼠和大鼠的RBP蛋白(mmrbp和rnrbp)被认为是在同一条进化分支上。但是，这个进化分支也包括那些只被64%引导复制方法所支持的小鼠、大鼠和兔子(ocrbp)的蛋白序列。这就表示了这三个蛋白的引导树有36%的可能性连接到其他一个蛋白质序列进化分类中，例如马的RBP序列(ecrbp)；或者这三个蛋白质序列成为其中某一种蛋白质(例如ocrbp)的进化分支中的一员。这样，引导检测方法就能通过在该数据集中采用重复取样分析的方法来为推论所得到树的拓扑结构给出一个评价准则

11.4 展望

分子系统发生分析是理解蛋白质（或核苷酸）序列进化方式和相互关系的基础工具。其中主要供分析的输出形式就是系统发生树，其事实上是对多重序列比对结果的一种图形化的表示。随着目前 DNA 和蛋白质序列数据的快速增长趋势，伴随着系统发生树的可视化冲击，使得系统发生分析变得越来越重要并且得到了较为广泛的应用。我们将会在第 12～第 16 章对系统发生的实例进行详细描述。

11.5 常见问题

基于分子序列数据的系统发生树的质量取决于分子序列的多重序列比对结果的优劣。为系统发生分析选择适合的模型也是非常必要的。最后，你对系统发生分析给出结果的理解也是非常重要的。每种重建系统发生树的方法都包含着很多的假设和对潜在缺点的包容。

11.6 网络资源

全球互联网上最好的系统发生资源的起点就是 Joe Felsenstein 的网站：http://evolution. genetics. washington. edu/phylip/software. html。该网站列举了 200 多个相关的链接，这 200 个链接分别按照系统发生方法、计算机平台、数据混合类型进行组织和分类。

11.7 问题讨论

使用基于距离的构建系统发生树的方法和基于字母特征的构建系统发生树的方法，是否对我们所关注的基因（或者蛋白质）家族给出完全不同的推论结果？

11.8 问题

序列比对结果的输入

① 登陆 NCBI 网站上的保守区域数据库（CCD）：http://www. ncbi. nlm. nih. gov/Structure/cdd/cdd. shtml。

② 在 CDD 主页右侧底部的位置你将会找到一个题名为“Find CDs”（寻找保守区）的选项；键入“lipocalins”，然后点击“Find CDs”选项。搜索的结果将是“pram00061”。点击打开这个序列家族。

③ 新打开的窗口将对 lipocalin 蛋白质家族做一个简要的介绍，并包括了 89 个有代表性的相似比对序列（值得注意的是，所得到的结果大于已知的 59 个 lipocalin 蛋白序列，但是 59 只是其在 PFAM 中已经登记的序列的数目）。

④ 选择 mFasta 程序，然后点击“View Alignment”（观察比对结果）。

⑤ 将 FASTA 程序比对的结果复制到一个简单的文本文件中。将比对结果的第一行中代表 *gi* 值的前 9 个字符替换成其他名字。例如将“gi | 809398”改成“btrbp”（bos taurus

retinol binding protein，牛维生素结合蛋白）将更加便于我们处理和理解比对结果。PAUP程序通常是不接受“/”和数字字符的；同时也应该去掉序列中的一致部分，这项工作可以通过去除被9个字母“consensus”（一致性）命名的一段蛋白起始序列来完成。

转化为便于阅读的文件形式（通过运行ReadSeq软件）

⑥ 在浏览器中打开一个新的窗口，并且登陆到ReadSeq的主页地址：http://searchlauncher.bcm.tmc.edu/seq-util/readseq.html。在任意搜索引擎中键入关键词“readseq”，都能很轻易地找到这个超链接地址。

⑦ 将FASTA多重序列比对产生的结果粘贴到相应的窗口中，选择PAUP/NEXUS输出格式，双击“Perform Conversion”（完成转换）按钮。

⑧ 得到输出结果并且复制到电脑的剪贴板中。将其复制成一简单的文本文件或记事本文件。从#NEXUS字符开始复制并且以“end”（结束）字符结尾。

输入到PAUP程序中

⑨ 打开PAUP程序。

⑩ 在文件菜单中选择“New”（新建）。

⑪ 然后在窗口中粘贴上你的比对序列结果。

⑫ 在文件菜单下，选择“Execute”（执行）。

⑬ 如果得到了一些错误的信息，可能需要去尝试做一些改变，例如：

a. 找到含有“format datatype=protein inter-leave missing=-”内容的那一行，并且把这一行中的连字号（-）改变成句点（.）。

b. 在编辑菜单中选择“Find”（查找）选项，将其中所有的“/”符号都改变成划线（-）。

⑭ 当“Execute”（执行）命令成功运行后，将会被告知：“数据矩阵共有10个分类（例如10个蛋白质）和163个字符（氨基酸）”。

用最大简约法（MP）来分析进化树

（1）启发式搜索

⑮ 在分析菜单中，选择“Parsimony”（简约法）。

⑯ 在分析菜单中，选择“启发式搜索”。点击“Search”选项并且使用默认设置寻找合适的进化树。

⑰ 一旦搜索结束以后，显示“启发式搜索状态”的小窗口上可能会出现一个可以点击关闭的按钮。这个程序记载了重排以后进化树的数目，并且尝试为这些树打分并从中找出最佳的进化树。

⑱ 纵观这些进化树。选择进化树菜单并且选择“打印进化树”。我们可能可以在众多的选项中看到一个是选择无根系统发生图还是进化分支图的选项。在系统发生显示中，进化树分支的长度是与氨基酸改变的数目成比例的，并且图中附带一标度。另外，在进化分支图中分支的长度和氨基酸的改变是不成比例的。进化分支图只是描述了物种和种群之间的进化关系。另外，如果发现得到同样最佳分数的进化树不只一棵，那么就可以选择是观察其中那棵特殊的进化树还是显示所有得到的进化树。

⑲ 你所得到的进化树的最长和最短进化分支上氨基酸改变的数目是多少？哪些分类（操作分类单元OUT）是连接在这些分支上的？

⑳ 如果你得到多于一棵的进化树，可以选择一棵最具倾向性的进化树。在进化树菜单

下，选择“计算支持率”。

(2) 进化树的评估　对随机系统发生树检测的准则是，比较从比对序列开始构建的已经得到的进化树的打分与符合 X 随机分布所产生的随机树分值分布之间的关系。

㉑ 根据长度为 100、1000 和 10000 的随机进化树的频率分布来评估进化树。

㉒ 在分析下拉菜单中选择“评估随机进化树”选项。改变随机树的数目并观察由其产生的平均值和方差值是怎样改变的。

㉓ 通过一个启发式搜索计算得到的进化树所得到的分值是否明显高于随机产生的进化树的分值？

㉔ 对最简约进化树重新运行启发式搜索并且监测其得分值。这次你的分值是多少？因为我们是对 10 个分类（操作分类单元：OTU）的蛋白质序列比对进行操作，所以还可以运行穷举搜索来寻找最大简约树。在分析下拉菜单中选择“评估随机进化树”选项，得到了怎样的分数？它与使用启发式搜索策略所得到的分数比较的结果如何？

引导测试是另外一种类型的重取样检测。这个测试的准则就是从原始序列比对的结果中随机取舍排列成柱状的相互独立的氨基酸序列数据。这个最新产生的数据集保持着与原始比对数据同样的大小。引导检测方法描述了一个特殊的分支设计例子的支持率。

㉕ 运行引导检测程序。下拉分析菜单并且选择“Bootstrap（引导检测）/Jacknife”选项，然后改变取样的数目。尝试对 1000 个和 10000 个复制单位的可替代取样进行分析并得出结果。

㉖ 在每一个测试之后，你可以查看这棵进化树→查看多棵进化树→打印用引导法得到的支持率→预览（可以将树的类型更改为“无根”的）。

㉗ 用 10000 的重复取样数目来分析进化树。根据你得到的引导测试值来判断你所得到的进化树中有多少强支持率（引导值＞70%）的分支，并且观察它们都是由哪些分类来构成的？

㉘ 你能否确定特殊分支上的组成部分是直系同源的，还是旁系同源的？需要哪些信息来支持这些推论？

改变输入文件

㉙ 重新回到 CDD 数据库并且调出 lipocalin 的“pfam00061”家族（参见前文所列的①～③点）

㉚ 通过选择“最先罗列出来的序列”选项来替代选择“10 个变化最为多样的序列”的默认选项（参见④）来作为序列比对的输出值（作为选择，你可以将输入的序列数目增加到 25 条）。

㉛ 重复上述练习并且观察其中的差异性。

㉜ 你可以定制输入序列的形式。选择“已经选择的序列”选项并且检查那些已经被检查的并罗列在这个菜单下的被选择序列（具体参见④）。

㉝ 利用基于距离的方法来分析进化树。这个方法是基于比较序列中成对差异性的数目和计算所得的序列之间的距离来构建系统发生树的。非常不幸的是，当某些突变位点随之发生了回复突变，并突变回原始的字符特征时，这些突变（特别是当你构建一棵基于 DNA 的系统发生树而不是基于蛋白质序列的系统发生树时）将可能被忽视。在分析菜单中选择“距离”选项。然后对该进化树进行启发式搜索。纵观该进化树，观察其中利用最大简约法构建的进化树分支上的分类（操作分类单元：OTU）是否合理。另外，可以对该进化树运行上

文所提到的随机测试或引导检测方法程序来评估其优劣。

㉞ 运用最大似然法来分析系统发生树。这是一个基于字符特征构建系统发生树的方法（例如最大简约法）。在这个例子中，我们是利用对所观察数据进行最大似然法估计来构建系统发生树的。PAUP 程序可能只允许利用比对后的 DNA 序列结果来构建系统发生树。

获得比对的 DNA 序列方法中的一个是找回 PopSet 数据库中的比对结果［访问 NCBI 的主站点并且选择 Entrez→单击 PopSet→browse（浏览）Popset 数据库，并从中找出最令你满意的比对结果］。将从 Popset 数据库中得到的 PHYLIP 格式的比对结果粘贴到 ReadSeq 程序中并且按上述过程运行。

一个利用最大似然法来分析蛋白质序列比对并构建其进化树的过程可以通过程序 Puzzle（http://www.tree-puzzle.de）来完成。

11.9 自测题

1. 根据分子时钟假说

(a) 所有的蛋白质都保持一个相同的恒定的进化速率；

(b) 所有的蛋白质进化速率都与化石记录相符合；

(c) 对于每一个给定的蛋白质，分子进化的速率是逐步减慢的，就像一个不准时的时钟一样；

(d) 对于每一个给定的蛋白质，其分子进化速率在所有的进化分支上大致是恒定的

2. 系统发生树的两个主要特征是

(a) 进化分支和进化节点；

(b) 树的拓扑结构和分支的长度；

(c) 进化分支和树根；

(d) 序列比对和引导检测方法

3. 下列哪一个是基于字母特征的系统发生分析的算法？

(a) 近邻连接算法；

(b) Kimura 算法；

(c) 最大似然算法；

(d) PAUP

4. 有两个最基本的构建系统发生树的方法：基于距离的和基于字母特征的方法。它们之间的基本差异性是

(a) 基于距离的方法含有一个用于定义分支顺序的中间数据矩阵，而基于字母特征的方法却没有；

(b) 基于距离的方法只应用于 DNA 数据，而基于字母特征的方法可以应用于 DNA 或者蛋白质数据；

(c) 基于距离的方法利用简约算法而基于字符特征的方法却没有；

(d) 基于距离的方法所产生的进化分支与进化时间长短成比例，而基于字母特征的方法却没有

5. 一个操作分类单元指

(a) 多重序列比对；

(b) 蛋白质序列；

(c) 进化分支；

(d) 进化节点

6. 对于一对给定的操作分类单元，下列哪一种说法是正确的？

(a) 其修正后的遗传距离大于或等于替换率；

(b) 替换率大于或等于其修正后的遗传距离

7. 转换的权重通常都比颠换大

(a) 正确；

(b) 错误

8. 构建和分析一棵系统发生树时产生的最常见的错误是

(a) 使用一个不好的多重序列比对结果

作为输入；

(b) 尝试在进化树上推测基因（或者蛋白质）的进化关系；

(c) 尝试在进化树上推测基因（或者蛋白质）间相互演化开的时间；

(d) 假设进化分支是单一发生的

9. Clustal X 程序可以使用或者脱离引导值来产生一棵用邻居连接算法生成的进化树

(a) 正确；

(b) 错误

10. 如你有 200 条滤过性病毒的 DNA 序列并且每条序列含有 500 个残基，你若想知道其中哪些残基对是相同的（或者是相似的），则使用下面哪一种方法是最有效的？

(a) BLAST 程序；

(b) 最大似然系统发生分析；

(c) 近邻连接系统发生分析；

(d) Popset 数据库

11.10 推荐读物

Mayr 的书从科学历史观的角度对进化学进行了出色的综述，并于 1982 年出版。关于分子进化学科有很多极好的教科书，包括 Graur 和 Li 在 2000 年以及 Li 在 1997 年出版的书籍。Maddison 和 Maddison 在 2000 年出版的书中描述的名为 MacClade 的软件包提供了一个有关系统发生分析的广泛而清楚的介绍。Hall 在 2001 年出版的书也很精彩和实用。国际生物技术信息中心提供了一个在线的介绍分子进化树相关信息的初级读本（从分类学和分子系统发生角度）(http://www. ncbi. nlm. nih. gov/About/primer/phylo. html)。

本章提供的一个详细彻底的有关系统发生分析的参考综述是 David Swofford（PAUP 程序的作者）和他的同事（Swofford 等）在 1996 年完成的。最后大力推荐 Thornton 和 DeSalle 在 2000 年发表的有关系统发生分析的综述。他描述了系统发生分析的准则（如基于字母特征和基于距离特征的方法）、直系同源和旁系同源的评估，以及介绍了比较基因组学中的系统发生分析。

参 考 文 献

Anfinsen, C. B. The Molecular Basis of Evolution. John Wiley and Sons, New York, 1959.

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., and Doolittle, W. F. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290**, 972-977 (2000).

Baxevanis, A. D., and Ouellette, B. F. *Bioinformatics*, 2nd ed. Wiley-Interscience, New York, 2001.

Cavalli-Sforza, L. L., and Edwards, A. W. F. Phylogenetic analysis: models and estimation procedures. *Amer. J. Hum. Genet.* **19**, 233-257 (1967).

Clote, P., and Backofen, R. *Computational Molecular Biology. An Introduction*. Wiley, New York, 2000.

Copley, R. R., Schultz, J., Ponting, C. P., and Bork, P. Protein families in multicellular organisms. *Curr. Opin. Struct. Biol.* **9**, 408-415 (1999).

Czelusniak, J., Goodman, M., Moncrief, N. D., and Kehoe, S. M. Maximum parsimony approach to construction of evolutionary trees from aligned homologous sequences. *Methods Enzymol.* **183**, 601-615 (1990).

Darwin, Charles. *The Origin of Species by Means of Natural Selection*. John Murray, London, 1859.

Dayhoff, M. O. *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, MD, 1978.

Dickerson, R. E. Sequence and structure homologies in bacterial and mammalian-type cytochromes. *J. Mol. Biol.* **57**, 1-15 (1971).

Dobzhansky, T. Nothing in biology makes sense except in the light of evolution. *Amer. Biol. Teacher* **35**, 125-129 (1973).

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. *Biological Sequence Analysis*. Cambridge University Press, Cambridge, 1998.

Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**, 368-376 (1981).

Felsenstein, J. Phylogenies from molecular sequences: Inference and reliability. *Annu. Rev. Genet.* **22**, 521-565 (1988).

Felsenstein, J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* **266**, 418-427 (1996).

Feng, D. F., and Doolittle, R. F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351-360 (1987).

Graur, D., and Li, W.-H. *Fundamentals of Molecular Evolution*, 2nd ed. Sinauer Associates, Sunderland, MA, 2000.

Haeckel, E. *The Riddle of the Universe*. Harper and Brothers, New York, 1900.

Hall, B. G. *Phylogenetic Trees Made Easy. A How-To for Molecular Biologists*. Sinauer Associates, Sunderland, MA, 2001.

Hein, J. Unified approach to alignment and phylogenies. *Methods Enzymol.* **183,** 626–645 (1990).

Hillis, D. M. Approaches for assessing phylogenetic accuracy. *Syst. Biol.* **44,** 3–16 (1995).

Hillis, D. M., and Bull, J. J. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biol.* **42,** 182–192 (1993).

Hillis, D. M., and Huelsenbeck, J. P. Signal, noise, and reliability in molecular phylogenetic analyses. *J. Hered.* **83,** 189–195 (1992).

Jukes, T. H., and Cantor, C. Evolution of protein molecules. In *Mammalian protein metabolism*, H. N. Munro and J. B. Allison, J. B. (eds.), Academic Press, New York, 1969 pp. 21–132.

Kimura, M. Evolutionary rate at the molecular level. *Nature* **217,** 624–626 (1968).

Kimura, M. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, 1983.

Li, W.-H. *Molecular Evolution*. Sinauer Associates, Sunderland, MA, 1997.

Maddison, D., and Maddison, W. *MacClade 4: Analysis of Phylogeny and Character Evolution*. Sinauer Associates, Sunderland, MA, 2000.

Margoliash, E. Primary structure and evolution of cytochrome c. *Proc. Natl. Acad. Sci. USA* **50,** 672–679 (1963).

Mayr, E. *The Growth of Biological Thought. Diversity, Evolution, and Inheritance*. Belknap Harvard, Cambridge, MA, 1982.

Moritz, C., and Hillis, D. M. Molecular systematics: Context and controversies. In D. M. Hillis, C. Moritz, and B. K. Mable (Eds.) *Molecular Systematics*. Sinauer Associates, Sunderland, MA, 1996, pp. 1–13.

Nei, M. *Molecular Evolutionary Genetics*. Columbia University Press, New York, 1987, pp. 39–63.

Nei, M. Phylogenetic analysis in molecular evolutionary genetics. *Annu. Rev. Genet.* **30,** 371–403 (1996).

Nuttall, G. H. F. Blood Immunity and Blood Relationship. Cambridge University Press, Cambridge, 1904.

Ohno, S. *Evolution by Gene Duplication*. Springer-Verlag, New York, 1970.

Philippe, H., and Laurent, J. How good are deep phylogenetic trees? *Curr. Opin. Genet. Dev.* **8,** 616–623 (1998).

Posada, D., and Crandall, K. A. MODELTEST: Testing the model of DNA substitution. *Bioinformatics* **14,** 817–818 (1998).

Saitou, N., and Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425 (1987).

Schmidt, H. A., Strimmer, K., Vingron, M., and von Haeseler, A. TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18,** 502–504 (2002).

Simpson, G. G. *The Meaning of Evolution: A Study of the History of Life and of Its Significance for Man*. Yale University Press, New Haven, 1952.

Strimmer, K., and von Haeseler, A. Quartet puzzling: A quartet maximun likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13,** 964–969 (1996).

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. Molecular systematics: Context and controversies. In D. M. Hillis, C. Moritz, and B. K. Mable (Eds.), *Molecular Systematics*, 2nd ed. Sinauer Associates, Sunderland, MA, 1996.

Thornton, J. W., and DeSalle, R. Gene family evolution and homology: Genomics meets phylogenetics. *Annu. Rev. Genomics Hum. Genet.* **1,** 41–73 (2000).

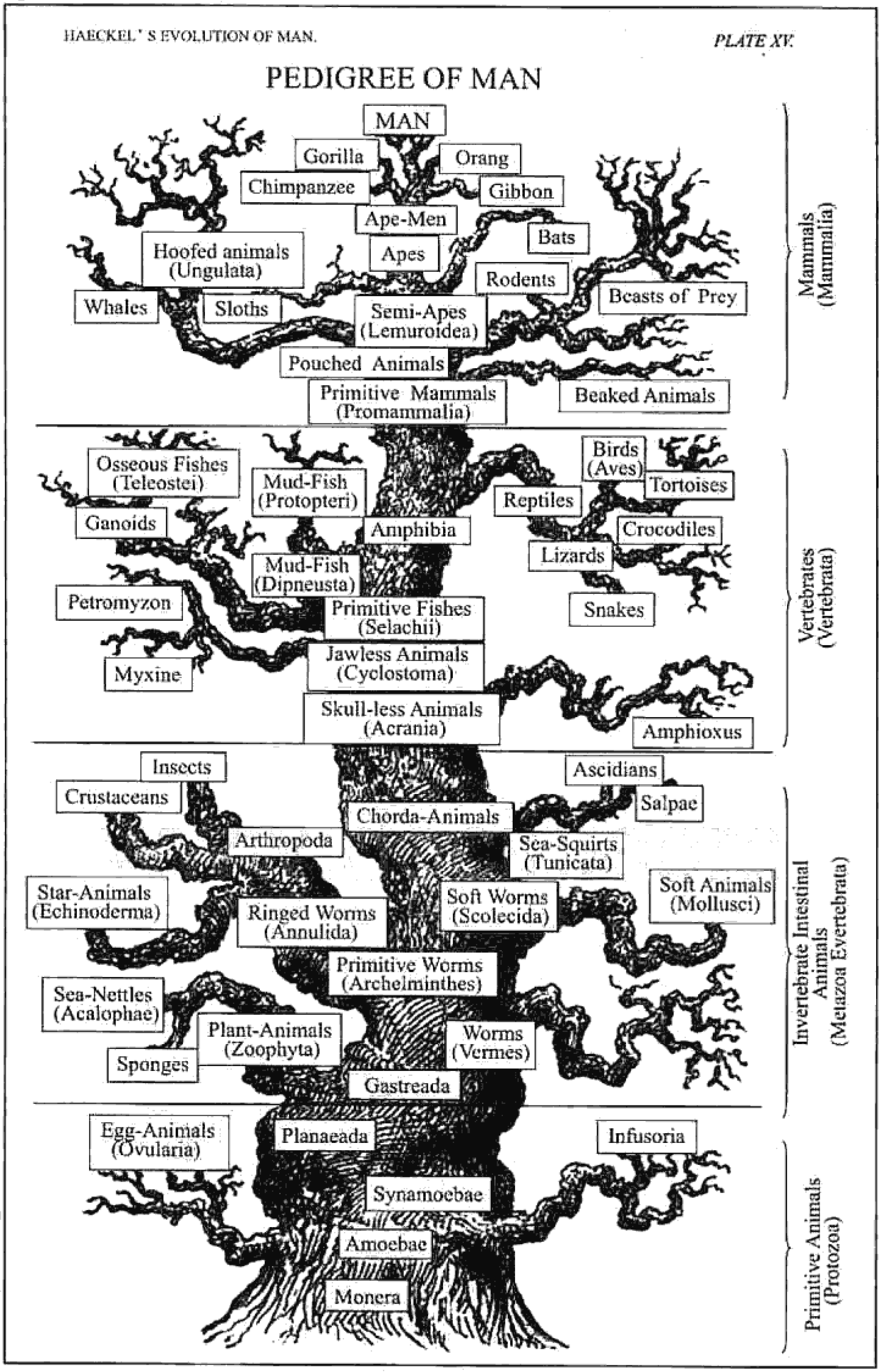
Wilson, E. O. *The Diversity Of Life*. W. W. Norton, New York, 1992.

Zuckerkandl, E., and Pauling, L. Molecular disease, evolution, and genic heterogeneity. In M. Kasha and B. Pullman (Eds.), *Horizons In Biochemistry*, Albert Szent-Gyorgyi Dedicatory Volume. Academic, New York, 1962.

第3篇

基因组分析

第 12 章　全基因组和系统发生树
第 13 章　已完成测序的基因组：病毒基因组
第 14 章　已完成测序的基因组：细菌和古细菌基因组
第 15 章　真核基因组：真菌基因组
第 16 章　真核基因组：从寄生生物到灵长类
第 17 章　人类基因组
第 18 章　人类疾病



此图来源于 Ernst Haeckel (1879) 的生命树，图上显示了哺乳动物（人类位于树的顶部显示来源于类人猿）、脊椎动物、无脊椎动物、原始动物位于底部，同时包括菌类（细菌）(使用经 Johns Hopkins 医学院医学历史研究所授权)

第 12 章

全基因组和系统发生树

同类的所有生物体之间的亲缘关系有时可以用一棵很大的生命之树来表示。并且我相信这种表示法很大程度上反映的是真实的情况。绿色的刚发芽的小枝可能代表的是现存的物种；那些之前长出来的枝条可能代表的是现存物种已经灭绝的祖先……进化树的树干分裂成为大的进化分支，然后它们又继续分裂成为更小的进化分支，就这样循环往复生长。这些较小的枝条与这棵树幼年时期刚发芽的小枝条没有什么不同；旧的和新的芽通过枝条分裂的形式相互联系在一起，正好代表了所有已灭绝的和现存的物种在分类中的类群隶属关系……从这棵树的最初生长开始，很多大的枝条都已经开始腐烂并且脱落，这些各种大小并且已经消失的枝条，代表了那些现在我们只能通过已经找到的物种化石才能了解的生物种群，并且已经没有现存物种的目、科和属等……芽长大以后又产生新的芽，只要生命力足够旺盛，它们就将不断分枝并取代那些较弱的枝条，若干代以后，我相信生命之树的发生生长过程也与此类似，地面上覆盖的都是那些已经死去或者断裂的枝条，而树的表面都是那些永不停止生长的美丽的分支。

——查尔斯·达尔文《物种起源》(1859)

12.1 引言

一个完整的基因组就是组成一个生物体的全部 DNA 的集合。每个独立生物个体的基因组包含了对其身份进行最终限定的一些特殊基因和其他一些 DNA 元素。基因组的大小范围变化很大，小的（如最小的病毒）只编码少于 10 个基因，而大的（如人类的真核生物）则包含数十亿个碱基对，编码数万个基因。

最近对几乎来自于所有进化分支的生物——包括多种病毒、原核生物、真菌、一种线虫、一种植物和人类基因组测序工作的结果，暗示现在我们正处在生物学发展历史中非常特殊的一个时期。类比一下可发现，目前的生物学发展情景类似于 19 世纪元素周期表完成时的大时代背景。当时人们已经逐渐清楚，各种化学元素可以排列成行列的形式，它将有助于预言各种化学元素的特性。慢慢地，解释各种化学元素特性的逻辑关系也相继出现了。但是，当人们真正完全掌握其中化学元素的意义和化学元素周期表内在的组织结构时，却又花去了整整一个世纪。

今天我们已经对数百个基因组进行了测序，现在需要找出一种能够解释它们相互之间组织结构以及功能的逻辑关系。这个过程可能要耗费数十年的时光，并且需要应用一系列的工

具，包括生物信息学方法、生物化学、遗传学和细胞生物学等的技术。

本章将介绍系统发生树的构建方法和全基因组的测序方法。我们将在接下去的每一章中对多种基因组的分析过程进行逐一的详细介绍，包括病毒（参见第13章）、原核生物（包括细菌和古细菌）（参见第14章）、真菌（包括酵母 *Saccharomyces cerevisiae*）（参见第15章）、从寄生虫到灵长类的一系列真核生物（参见第16章）的基因组，最后是人类基因组（参见第17章、第18章）。关于系统发生树一些关键术语的定义请参见表12-1。

表 12-1 系统发生树的相关术语介绍

名字	同义词	定义
archaea（古细菌，单数形式 archaeon）	archaebacteria	生命形式的三个"原界（urkingdoms）"或"域（domains）"之一
bacteria（细菌）	eubacteria；monera（旧称）	生命形式的三个"原界（urkingdoms）"或"域（domains）"之一；以缺少核膜为主要特征的单细胞生物
eukaryotes（真核生物）	eucarya	生命形式的三个"原界（urkingdoms）"或"域（domains）"之一；具有以核膜为主要特征的细胞
microbe（微生物，细菌）		指能导致人类疾病的微生物（Microorganism）；包括原生动物和真菌等真核生物以及细菌
microorganism（微生物，微小动植物）		显微尺度的单细胞生命形式，包括细菌、古细菌和一些真核生物
progenote（始祖生命）	最近的共同祖先	远古的单细胞生命形式，三个域的所有生物的共同祖先
prokaryotes（原核生物）	procayotes；以前是细菌的同义词	缺少核膜的生物体，细菌和古细菌

注："名字"指的是本书所采用的名字，请参考 Woese 等（1990）。

12.2 系统分类学简史

种是指在通常条件下只在同种之间相互繁殖的一群相似的生物体。一个属可能包含一个到几百个种。

自有文字记载的历史以来，哲学家和科学家们之间就不断发生有关地球上生命形式多样性问题的争论（Mayr 于 1982 年提出）。亚里士多德（Aristotle，公元前 384～前 322 年）同时也是一位积极的生物学家，他在其有关动物学的著作中描述了 500 多种物种。他虽然没有给生命形式建立一种普适的分类方案，但是在其"Historia animalium"（《动物史》）的著作中，的确是将动物描述成为"有血"或者"无血"的两类[最后，拉马克（Lamarck，1744～1829）将这种分类改进为"有脊椎"和"无脊椎"两类]。亚里士多德将动物按属和种的分类方式也成为我们今天所使用的分类学系统的分类原型。

这种将每个生物体按属和种的方式来命名的双命名体系的最主要提倡者是瑞典博物学家林奈（Carl Linnaeus，1707～1778）。林奈还提出了动物界、植物界和矿物界三界的概念；他的分级系统所采用的四个等级分别是纲、目、属和种。海克尔（Ernst Haeckel，1834～1919）进一步扩大了这个系统，一共描述了将近四千多个新物种。他将生命形式描述为从纯粹的复合分子到植物和动物的连续统一体，而将原核生物（Moner）描述为无定形的生命群簇。其中原核生物（monera）后来被人们命名为细菌（bacteria）。到了 20 世纪 60 年代末，海克尔、科普兰（Copeland）、魏泰克（Whittaker）以及其他许多研究者的工作产生了最终的标准生命五界系统：动物界、植物界、单细胞原生生物界、真菌和原核生物界。本章首页的插图就是从 1879 年海克尔的书上摘下来的一棵系统发生树的示例。在五界系统形成的同时，查顿（Edouard Chatton，1937）提出了原核生物（无细胞核的细菌）和真核生物（由有核细胞构成的生物体）之间的区别。

伍斯（Carl Woese）和他的同事们（Fox 等，1980；Woese，1998；Woese 等，1990）在 20 世纪 70 年代和 20 世纪 80 年代分别重建了系统发生树。他们研究了一组单细胞的生命形式，其因为缺少细胞核而被认为是细菌的原核生物。研究人员对小亚基核糖体 RNA（SSU rRNA）进行了测序和系统发生分析，结果发现古细菌在系统进化分析上较细菌更接近于真核生物。对在所有已知生命形式中都存在的小亚基核糖体 RNA（SSU rRNA）序列的系统发生分析可以提供构建系统发生树的一种版本（图 12-1）。图中一共存在三个主要的进化分支。因为这棵进化树的确切根部仍然未知，而分支最深的细菌和古细菌都属于嗜温性生物，因而可以推测生命形式或许是来源于一个高温的自然环境。

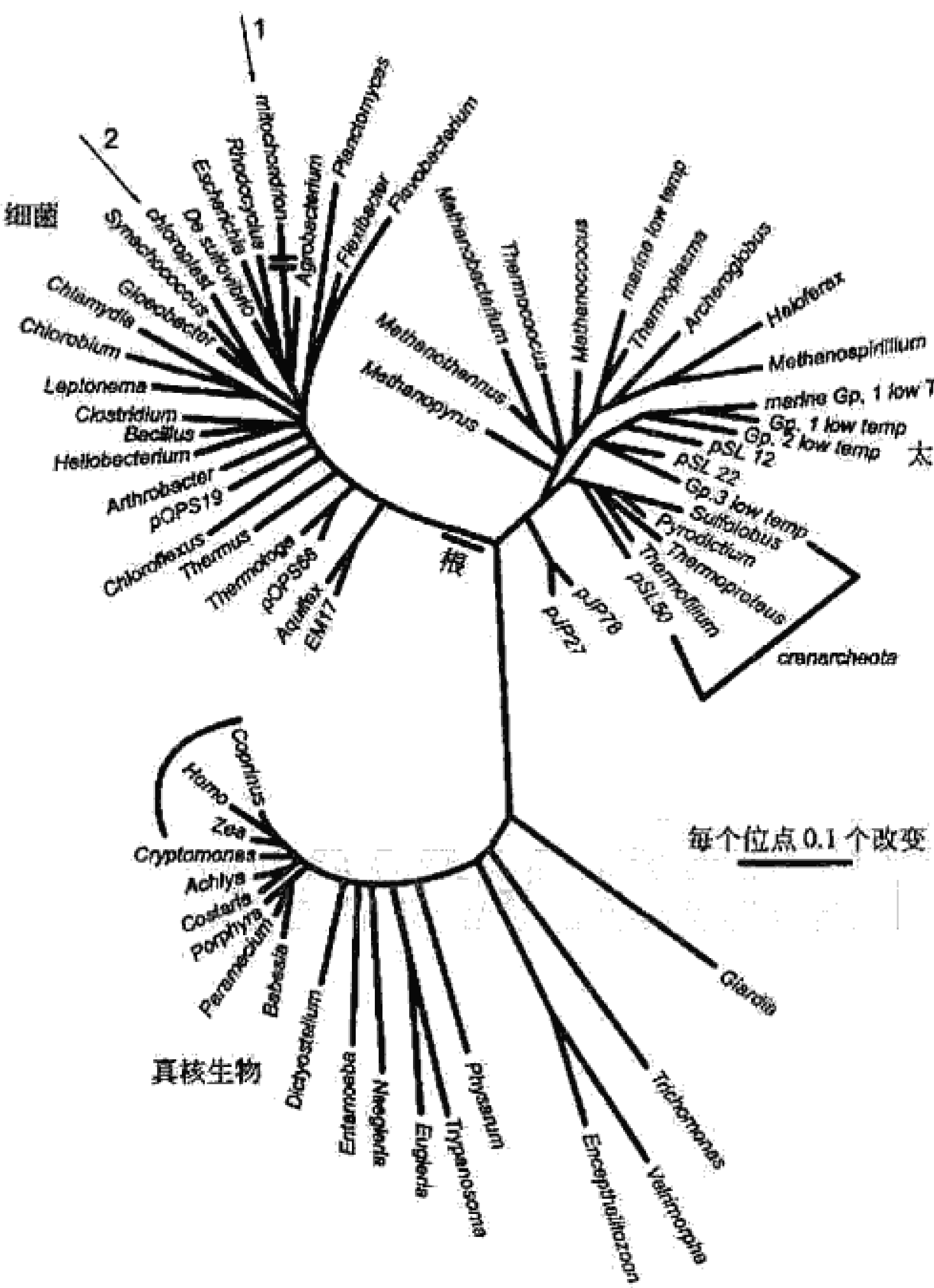


图 12-1 总的系统发生树，基于小亚基 rRNA 序列的系统发生分析（Pace 于 2001 年修改）。目前普遍认为生命形式源于至少 38 亿年以前的厌氧环境中。最原始的生命形式（始祖生命）已经显示出了明显的生命特征（自我复制和进化）。数据授权使用

病毒不满足活生物体的定义，因此它们不出现在这棵系统发生树中。虽然它们可以复制并且进化，但是病毒只能通过霸占活生物体的细胞而生存。

12.3 地球上生命形式的生物发展史

我们前面看到的系统发生树（图 12-1）包含了对地球上生物发展史的全新解释。所有的生命形式有一个共同的起源，并且构成了系统发生树的某个部分。物种的平均半衰期为 100 万年至 1000 万年，而且曾经生存在地球上的 99%以上的物种现在都已经灭绝了（Wilson，1992）。原则上，世界上只有一棵系统发生树能够准确地描述出物种的进化过程。系统发生学的目标就是试图

推出这一棵能够正确反映物种进化以及基因和蛋白质序列同源关系的系统发生树。系统发生学的另一个目标是推断出离现在时间距离最短的、共有一个祖先的生物体之间的分化时间。

图 12-2 显示了生物发展史的总览。最初的生命迹象始于大约 40 亿年前，即地球形成仅 0.5 亿年之后。最初的生命形式是以 RNA（而不是 DNA 或蛋白质；Joyce，2002）为中心

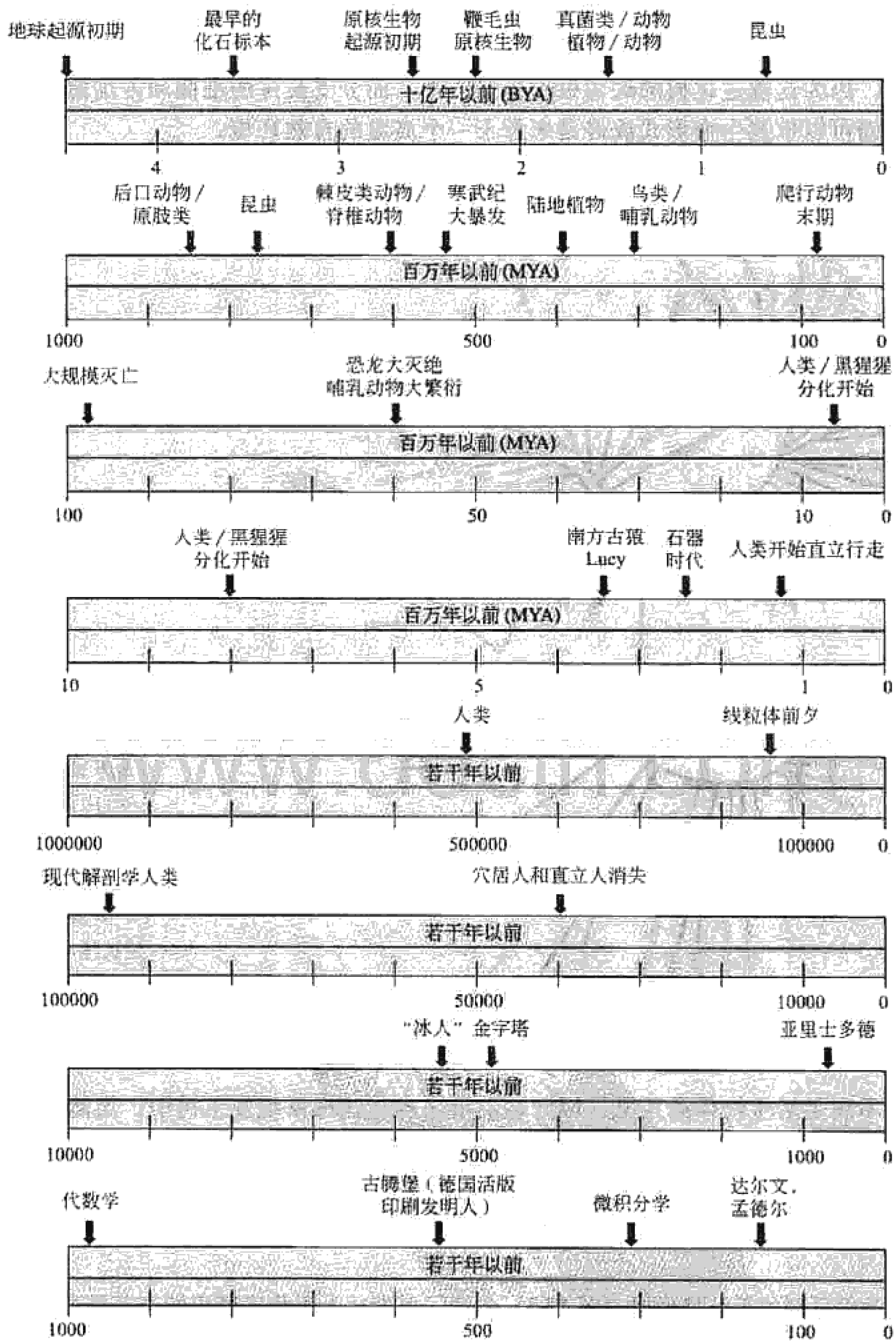


图 12-2 地球的生物发展史。来源包括 Kumar 和 Hedges（1998），Hedges 等（2001），Benton 和 Ayala（2003）

的。在早期的进化过程中，地球的大气环境为厌氧性，这种最初的生命形式据推测应当是一种单细胞的原核生物。反映最早生命迹象的化石据测算存在于距今 38 亿年前。通过谱系分化并演变成为现代细菌和现代古细菌之前离现在时间距离最短的共同祖先，其极可能是嗜高温的生物。其根据来源于进化树（图 12-1）中最底层分支上的物种，如细菌 *Aquifex* 和嗜高温的古细菌（请参见第 14 章中相关描述）。真核生物出现在距今大约 20 亿～30 亿年前，而且其在 10 亿年前始终保持着单细胞的形式，在距今大约 15 亿年前分化出植物、动物以及细菌，并且产生了多细胞生物。最近的 10 亿年间相继大量出现了复杂的多细胞生物的进化。距今 5 亿 5000 万年前所谓的寒武纪暴发，见证了动物生命形式多样性的惊人增长。在过去的 2 亿 5000 万年中，陆地聚合成了一个巨大的大陆，名为 Pangaea（泛大陆，参见图 12-3）。当 Pangaea 再度分裂成南北两个大洲（Laurasia 和 Gondwana）时，就给生命的繁殖设立了自然的障碍，影响了往后生物的进化。到了大约距今 6000 万年前，恐龙灭绝，哺乳动物的统治开始。

多细胞生物曾多次独立地进化。几十亿年前多种多细胞细菌的进化，对逃避捕食者有选择性的益处（Kaiser，2001）。

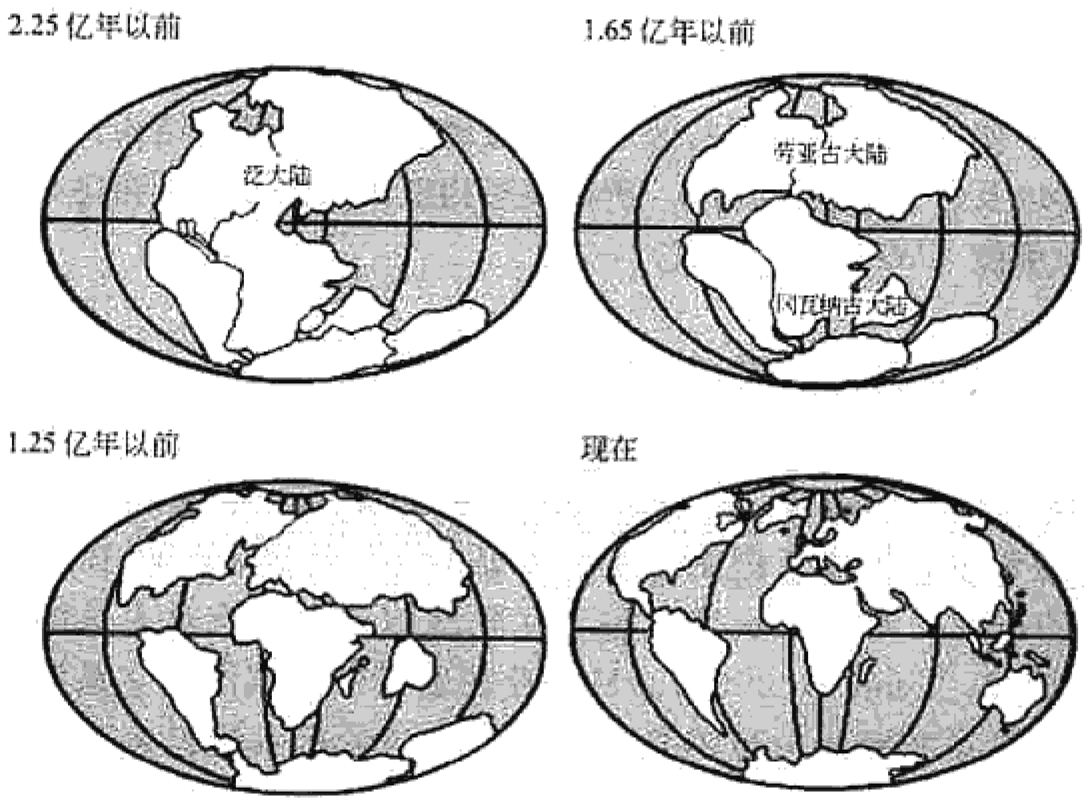


图 12-3　从距今 2.25 亿年前到现在地球的地理变化。起初，地球只有一个超级大洲（Pangaea）；到了距今 1.65 亿年前，Pangaea 分裂成为 Laurasia（劳亚古大陆，现代的亚洲和北美洲）和 Gondwana（冈瓦纳古大陆，现代的非洲和南美洲）；并且在距今 1.25 亿年前 Laurasia 和 Gondwana 各自开始分裂并最终形成了现在的七大洲的格局

12.4　系统发生树的分子序列基础

在过去的几个世纪里，提出的有关系统发生树的基本模型主要是形态学上的。例如林奈将动物分为六个纲（哺乳动物纲、鸟纲、鱼纲、昆虫纲、爬行动物纲和蠕虫纲），然后再按着哺乳动物的牙齿特征、鱼类的鳍类型以及昆虫的翅膀等特征细分下去。早期的显微镜观察发现了细菌没有细胞核，这使得细菌和其他四个界得以根本地区分开来。细菌可以按照其生物化学的特性（如 Albert Jan Kluyver 提出的，1888～1956）进行分类，而且从形态学角度细菌还可以分为几个主要的组。然而，这样的准则无法满足微生物纷繁复杂的物种多样性需求。因此，物理的准则无法成为将古细菌作为一个独立的生命分支区别开来的理由。

欧洲小亚基核糖体 RNA 数据库（Van de Peer 等，2000；Wuyts 等，2002）（http://rrna.uia.ac.be/ssu/或 http://oberon.rug.ac.be:8080/rRNA/）目前有超过 20000 个小亚基核糖体 RNA 的序列（2001 年 9 月），大约有 600 个古细菌物种，12000 个细菌和 6500 个真核生物。你可以获取这个数据库的序列并生成相应的系统发生树或者在线查看这些进化树。

分子序列数据的出现使我们对生命形式的研究得以更进一步。这些数据最初产生于 20 世纪 50 年代和 20 世纪 60 年代，到 1978 年，戴霍夫（Dayhoff）的图集使用了几百个蛋白质序列作为其 PAM 矩阵的基础（详情参见第 3 章中的相关描述）。最近几年 DNA 的序列信息得以飞速的增长，基于表形特征和基于基因序列的系统发生分析现在都已成为可能。使用最广泛的序列是小亚基核糖体 RNA 分子，它存在于几乎所有现存的生命形式中。小亚基核糖体 RNA 缓慢的进化速度以及合适的大小使得它们非常适合于做系统发生的分析。基因组测序正在改变进化领域的研究，它将为系统发生树提供数以千计的 DNA 和蛋白质序列。

已经有超过 100 个原核基因组被测序（详情参见下文及第 14 章中的相关描述）。现在我们已经开始关注水平基因转移（详情参见第 14 章中的相关描述），这种现象指的是一种物种并不是从其血统关系而是从横向（或者侧向）的其他无关物种中获取某个基因的现象。因此，基因是可以在物种之间发生交换的（Eisen，2000）。其结果是，在分子系统发生中使用不同的个别基因常常会导致完全不同的树结构。正是因为生物界存在着水平基因转移以及基因丢失的现象，所以要想构造出反映地球生命进化过程的系统发生树也许永远都不太可能（Wolf 等，2002）。

12.5 生物信息学在系统分类学中的角色

生物信息学研究领域主要涉及到使用计算机算法和计算机数据库来阐述生物学的原理。生物信息学在系统分类学中的研究范围主要包括对基因、蛋白质和细胞在各种生物体环境中系统发生树结构的研究。有人建议建立基于互联网的分类系统对大量的生物数据进行归类（Blackmore，2002；Pennisi，2001）。目前存在着若干个试图建立系统发生树的计划。也有人建议虽然基于互联网的设想不错，但是目前的发生系统已经足够了：动物、植物或者其他的生物样本都已经在国际公约制定的方针下进行了采集、命名和研究（Knapp 等，2002）。

生物多样性公约组织（The Convention on Biological Diversity，网址：http://www.biodiv.org/）和全球生物多样性信息机构（Global BiodiversityInformation Facility，网址：http://www.gbif.net/）是其中两个关注全球生物多样性问题的组织。系统发生树计划位于 http://www.panspermia.org/tree.htm，系统发生树网络计划（David R. Maddison 创建）位于 http://tolweb.org/tree/phylogeny.html。

12.6 基因组测序计划

20 世纪 70 年代 DNA 测序技术的出现，包括桑格（Fred Sanger）的双脱氧核苷酸方法，使得大规模的测序计划得以实施。本章将详细介绍基因组测序计划的简要历史，包括 1995 年首个独立生活的生物体流感嗜血杆菌的基因组序列测序的完成。到 2001

年，两个小组报告了人类基因组的草图。目前全基因组测序方面最为引人注目的成果就是每年所收集序列数据的急剧增长（参见图 2-1）。随着基因组 DNA 的测序能力从最初的几百万个核苷酸残基发展到现在的几十亿个，科学界正面临着前所未有的机遇和挑战。

近几年的主题（Webb，2001）：

① 产生的序列数据量保持持续的急剧增长。

② 许多基因组，即使是那些未完成的基因组序列数据——即基因组中可能含有许多间隙和测序错误的序列——都可以立即被取得并被科学界所使用。而已完成的序列（定义见下文）比未完成的序列则提供了更好的基因组特征描述和注释。

③ 比较基因组分析目前用于解决诸如识别人类和小鼠的蛋白质编码基因或比较病原体的有毒和无毒菌株区别之类的问题（Koonin，2001）。通过保守 DNA 成分的分析，比较基因组分析还可用于划定基因的调控区域以及研究物种的进化历史。

12.7 基因组测序计划的简要年表

12.7.1 总览：1977 年～现在

数十个基因组测序计划很快就完成了，可以预期以后的进程会进一步加快。一个年表总览将有助于我们建立这些事件的基本框架。1995 年当第一个细菌基因组测序刚刚完成时，还只有很少量的其他基因组序列可供比较。而现在已经有 1000 多个全基因组可用了（包括细胞器的全基因组）。

在本章以及后续的章节中，我们将介绍 3 个有关基因组研究的主要网络资源：EBI、TIGR 和 NCBI（这些以及其他的站点已列在表 12-2 中）。我们将从 NCBI 的 Entrez 的基因组部分（图 12-4）开始。点击左边侧栏中的“Prominent Organisms（主要的生物体）”可以看到真核生物、细菌和古细菌基因组的网站总览（图 12-5）。

表 12-2 全基因组计划中涉及的互联网资源

资源	描述	URL
EBI	欧洲生物信息学研究所(European Bioinformatics Institute)	http://www.ebi.ac.uk/genomes/
GNN	基因组新闻网(Genome News Network)	http://gnn.tigr.org/sequenced_genomes/genome_guide_pl.shtml
GOLD	基因组在线数据库(Genomes Online Database)	http://ergo.integratedgenomics.com/GOLD/
Infobiogen	微生物全基因组	http://www.infobiogen.fr/doc/data/complete_genome.html
NCBI	美国国家生物技术信息中心(National Center for Biotechnology Information)的 Entrez 系统	http://www.ncbi.nlm.nih.gov/Entrez/
PEDANT	蛋白质提取、描述和分析工具(Protein Extraction, Description and Analysis Tool)	http://pedant.gsf.ed/
TIGR	基因组研究所(The Institute for Genomic Research)	http://www.tigr.org/tdb

注：这里的网站大多包含到其他更多相关网站的链接。

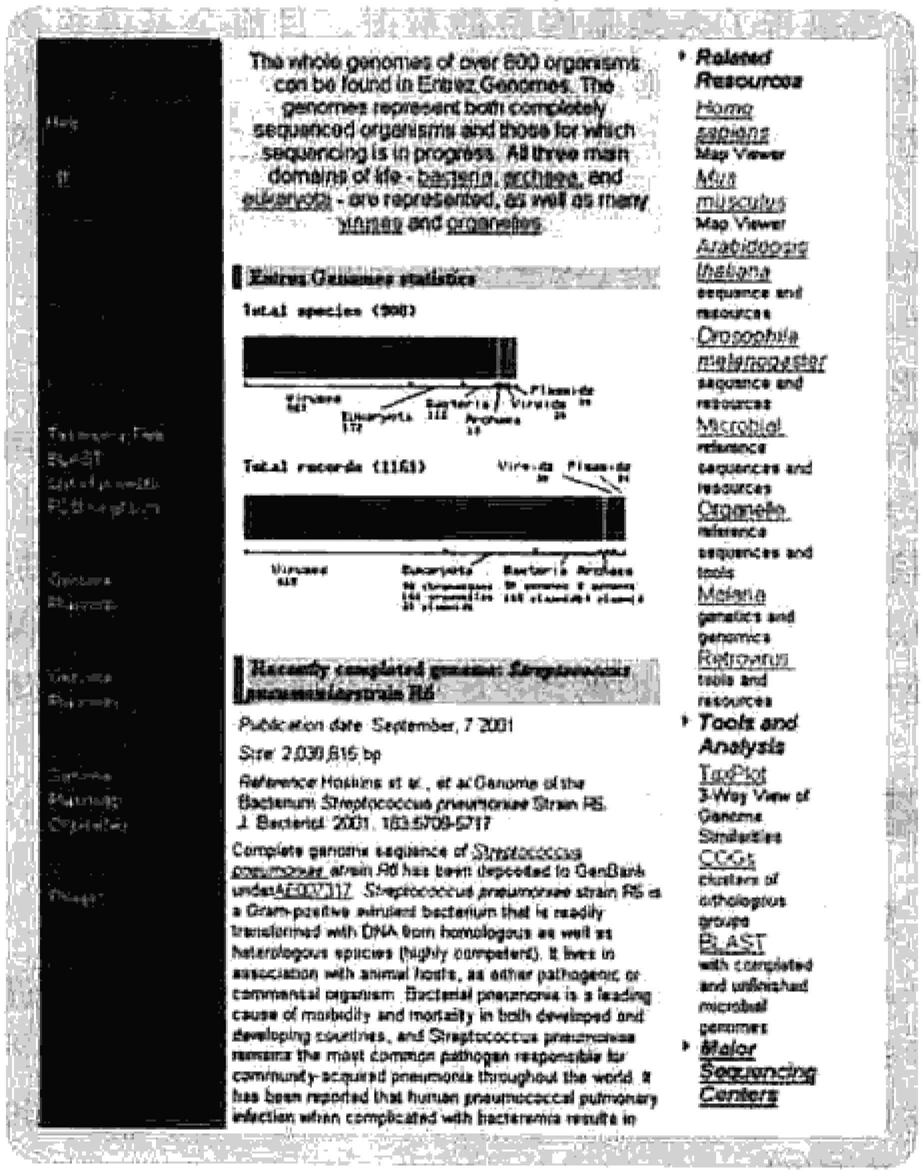


图 12-4 NCBI 的 Entrez 基因组站点包含超过100000个物种的分子序列数据的链接（左边侧栏）。除了几个基因组的分析工具外，还有一些与选中的生物相关的资源信息（右边侧栏）。核苷酸序列覆盖了超过900个的基因组。本图来自于 http://www.ncbi.nlm.nih.gov/Entrez/

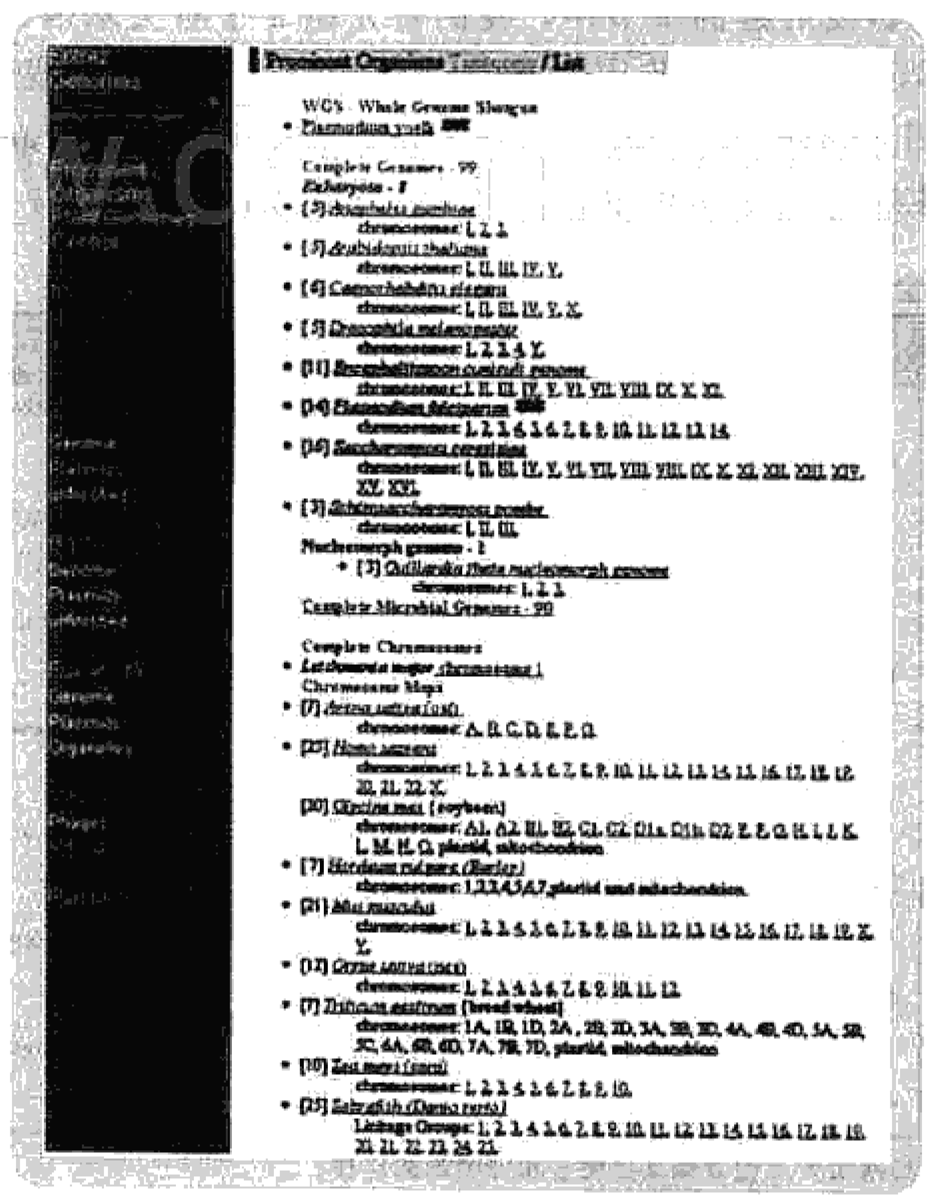


图 12-5 Entrez 基因组的 Prominent Organisms（主要物种）页面中列出了已经编码基因组的链接，包括微生物基因组和选中的真核生物的染色体图

12.7.2　第一个病毒基因组（1977 年）

桑格和其同事们对 ϕX174（Sanger 等，1977）基因组进行了测序。他们发展了多种 DNA 的测序技术，包括双脱氧核苷酸链终止测序法。ϕX174 包含了 5386 个碱基对，一共编码 11 个基因（参见 GenBank J02482）。图 12-6 是这个病毒基因组一部分的图形描述。最令人感到奇怪的结果是那些之前没有被预料到的在不同阅读框中转录的重叠基因的存在。

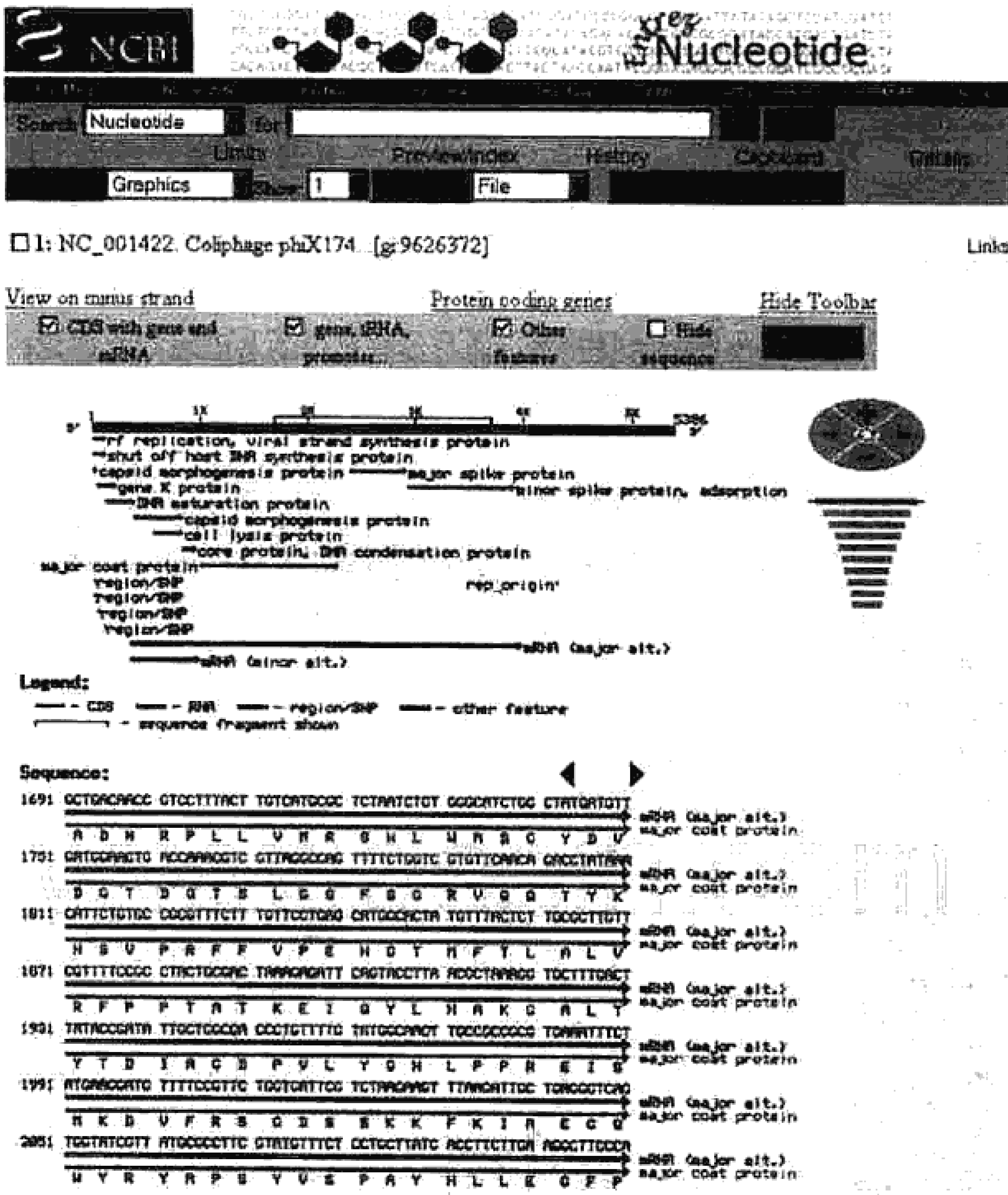


图 12-6　Entrez 中的 ϕX174 核苷酸记录的一部分。这个格式是以图形显示方式查看编号为 J02482 的条目所得到的。它提供了系统预测的开放式阅读框（ORFs）的总览。在底部显示的是部分核苷酸序列以及与其对应的预测蛋白质

在双脱氧核苷酸链测序法中，模板 DNA 在放射性核苷酸［如（^{32}P）dATP，用作观察新合成的 DNA 有义链的跟踪器］、四种核苷酸（dATP、dGTP、dCTP、dTTP）以及少量双脱氧核苷酸［ddATP、ddGTP、ddCTP、ddTTP（dideoxythymidine triphosphate，双脱氧胸苷三磷酸）中的一种］存在的条件下被聚合酶复制。双脱氧核苷酸使链的复制终止，导致生成的产物在链 5′端的核苷酸基团是相同的，但是在链的 3′端不相同。这些产物通过电泳聚集到聚丙烯酰胺凝胶上，就可以推断出其碱基的组成成分了。

12.7.3 第一个真核细胞器基因组（1981 年）

第一个被测序的细胞器基因组是人类的线粒体（Anderson 等，1981）。这个基因组的特点是只包含极少量的非编码 DNA。绝大多数多细胞动物的线粒体基因组大约都是 15～20kb 的环状基因组。人类的线粒体基因组大约为 16568bp，一共编码 13 个蛋白质、2 个核糖体 RNA 和 22 个转录 RNA。这些数据可以从 NCBI Entrez 基因组网站上获取（图 12-7）。网页上环形的图表是可以点击进入的，可以访问其上的每一个基因（图 12-8）。因此，所有线粒体基因的 DNA 和对应的蛋白质序列都可以轻易地获得。

> 更多的线粒体基因组的信息可从 http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/mitostat.html 获取。

今天，大约已经有超过 400 个线粒体的全基因组序列。其中几个列在表 12-3 中。这个表同时也列出了一些异常大的个例。最大的已经测序的线粒体基因组是拟南芥（*Arabidopsis thaliana*）(367kb）的线粒体基因组，但是据报告还有多种植物含有甚至更大的线粒体基因组。因此，线粒体基因组的多样性非常显著（Lang 等，1999）。虽然真核生物中的线粒体可能与细胞核中的基因组同时进化，但是分子系统发生方法通常认为线粒体是一种内共生的 α 蛋白细菌的后代（Lang 等，1999）。

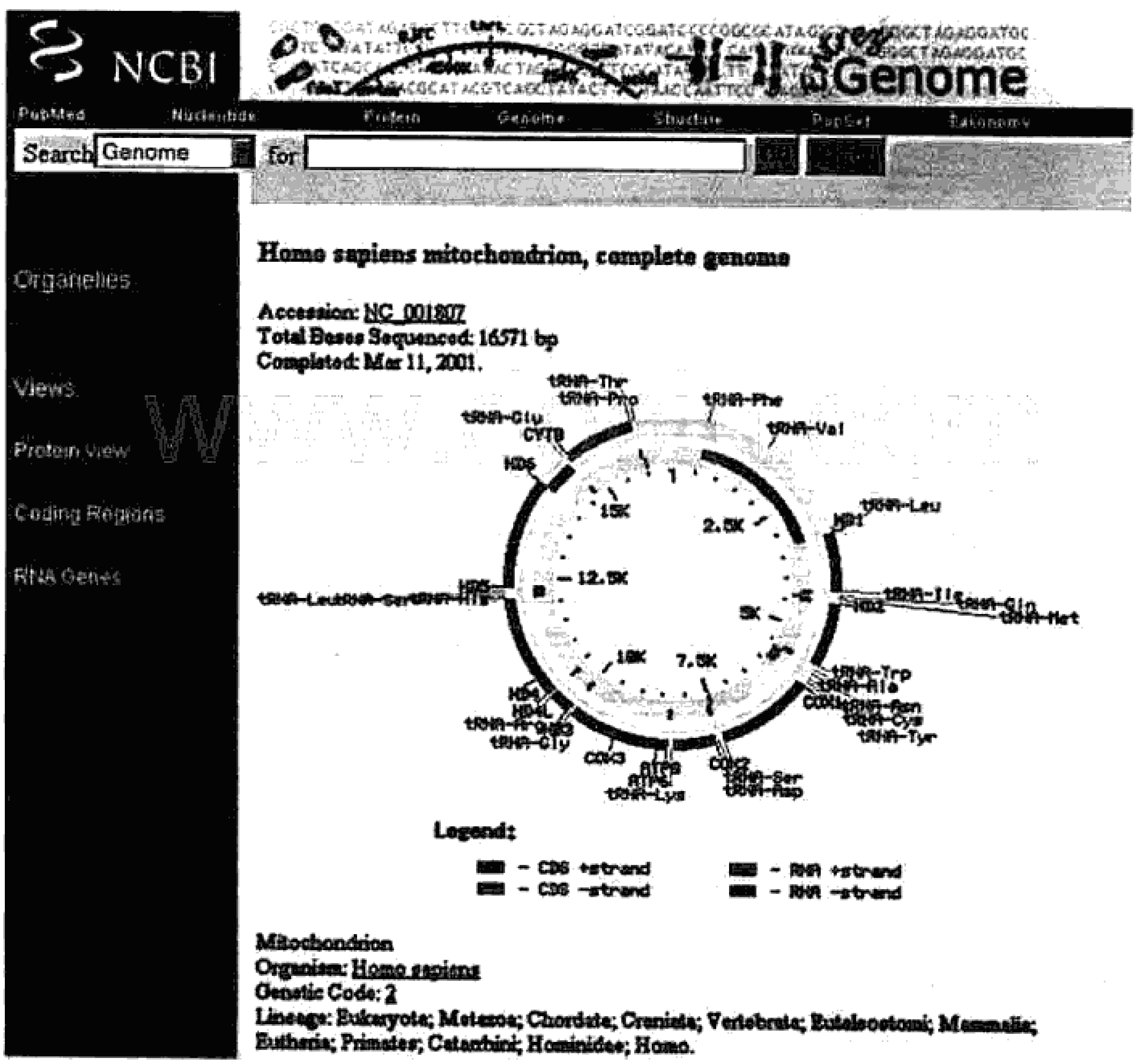
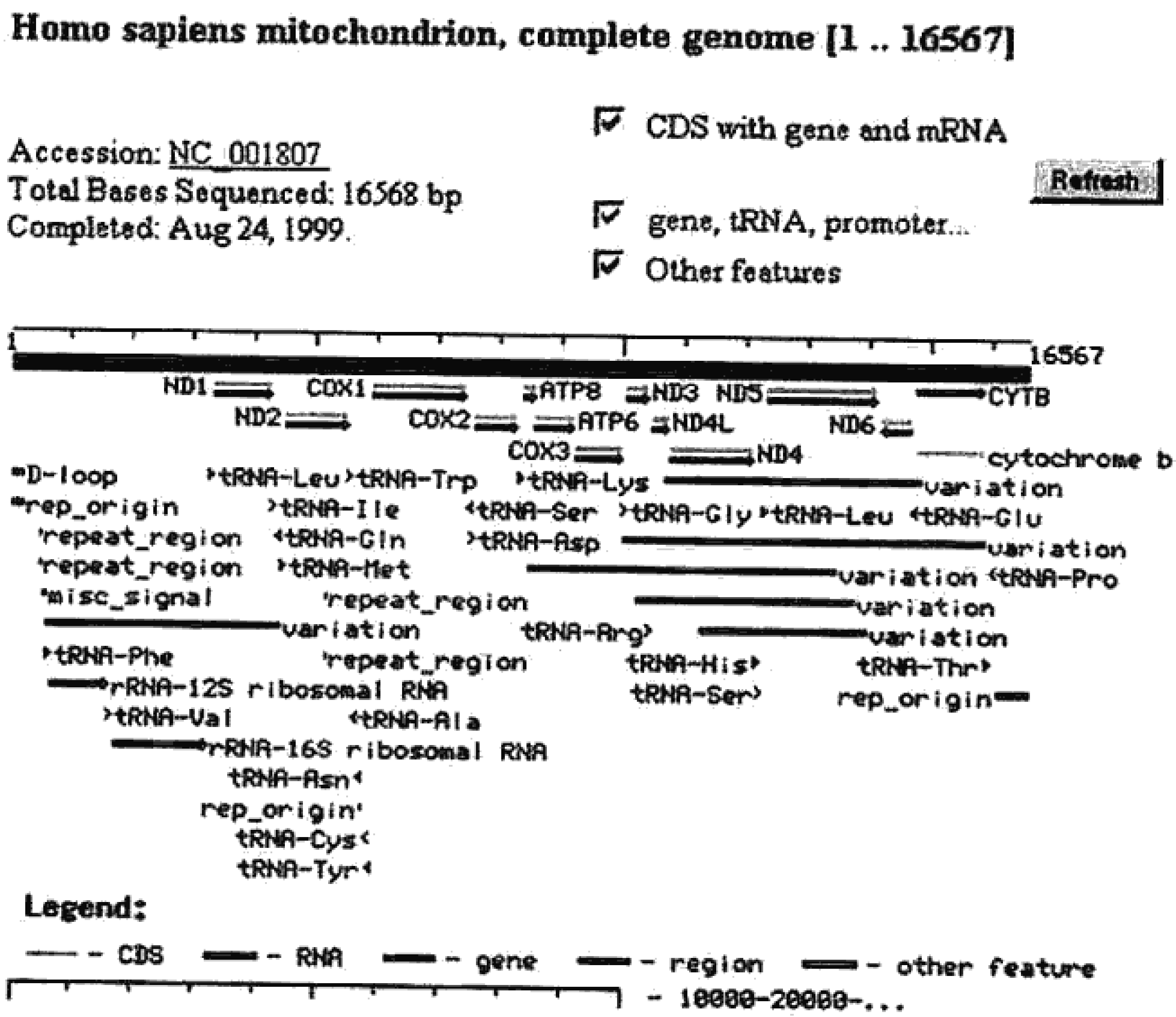


图 12-7 人类线粒体基因组的 NCBI 入口（NC_001807；16568bp）。点击主图［参见图 12-8(a)］或点击左边侧栏中的“Coding Regions（编码区域）”链接［参见图 12-8(b)］可获得相关序列的更多信息

(a)



(b)

Homo sapiens mitochondrion, complete genome

Save the report below in Table format.

◆ - GenBank record including protein ◆ - DNA region in flatfile format ◆ - DNA and protein in FASTA format

	Location	Strand	Length	PID	Gene	Synonym	Product
◆ ◆ ◆	3306..4262	+	319	5835388	ND1		NADH dehydrogenase subunit 1
◆ ◆ ◆	4469..5512	+	348	5835389	ND2		NADH dehydrogenase subunit 2
◆ ◆ ◆	5903..7444	+	514	5835390	COX1		cytochrome c oxidase subunit I
◆ ◆ ◆	7585..8268	+	228	5835391	COX2		cytochrome c oxidase subunit II
◆ ◆ ◆	8365..8571	+	69	5835392	ATP8		ATP synthase FO subunit 8
◆ ◆ ◆	8526..9206	+	227	5835393	ATP6		ATP synthase FO subunit 6
◆ ◆ ◆	9206..9988	+	261	5835394	COX3		cytochrome c oxidase subunit III
◆ ◆ ◆	10058..10405	+	116	5835395	ND3		NADH dehydrogenase subunit 3
◆ ◆ ◆	10469..10765	+	99	5835396	ND4L		NADH dehydrogenase subunit 4L
◆ ◆ ◆	10759..12138	+	460	5835397	ND4		NADH dehydrogenase subunit 4
◆ ◆ ◆	12336..14147	+	604	5835398	ND5		NADH dehydrogenase subunit 5
◆ ◆ ◆	14148..14672	-	175	5835399	ND6		NADH dehydrogenase subunit 6
◆ ◆ ◆	14746..15880	+	378	6137797	CYTB		cytochrome b

图 12-8 (a) 点击指定的基因(例如 *COX1*；参见图 12-7)可以获得人类线粒体基因组(NC_001807)的详细信息；(b) 线粒体基因组页面左边侧栏中的链接给出了所有线粒体的蛋白质编码基因的列表

12.7.4 第一个叶绿体基因组(1986 年)

在地钱(*Marchantia polymorpha*)的基因组公布之后(Ohyama 等，1986)，又公布了第一个叶绿体基因组的报告(Nicotiana tabacum；Shinozaki 等，1986)。大多数植物的叶绿体基因组大小为 120000～

我们将在第 16 章的植物部分讨论叶绿体和其他质体。

表 12-3 按大小排列的线粒体基因组选集

界	物种	编号	大小/bp
真核生物	*Plasmodium falciparum*(疟原虫)	NC_002375	5967
后生动物(两侧对称的动物)	*Caenorhabditis elegans*(蠕虫)	NC_001328	13794
植物(绿藻)	*Chlamydomonas reinhardtii*(绿藻)	NC_001638	15758
后生动物(两侧对称的动物)	*Mus musculus*(小家鼠)	NC_001569	16295
	Pan troglodytes(黑猩猩)	NC_001643	16554
后生动物	*Homo sapiens*(人类)	NC_001807	16568
后生动物(腔肠动物)	*Metridium senile*(海葵)	NC_000933	17443
后生动物(两侧对称的动物)	*Drosophila melanogaster*(果蝇)	NC_001709	19517
真菌(子囊菌)	*Schizosaccharomyces pombe*(粟酒裂殖酵母)	NC_001326	19431
	Candida albicans(白色念珠菌)	NC_002653	40420
真核生物(*Stramenopiles*)	*Pylaiella littoralis*(蓝藻)	NC_003055	58507
真菌(子囊菌)	*Rhizophydium* sp. 136(后生单鞭毛游孢子的子囊菌 136)	NC_003053	68834
真核生物	*Reclinomonas americana*(原生生物)	NC_001823	69034
真菌(子囊菌)	*Saccharomyces cerevisiae*(啤酒酵母)	NC_001224	85779
植物(*Streptophyta*)	*Arabidopsis thaliana*(拟南芥)	NC_001284	366923
植物	*Zea mays*(玉米)	Fauron 等(1995)	700000
	Cucumis melo(甜瓜)	Lilly 和 Havey(2001)	2400000

注：到 2003 年 6 月为止，357 个多细胞动物的细胞器基因组已经完成测序，其中包括 23 种真菌、10 种植物和 23 种其他真核生物（http://www.ncbi.nlm.nih.gov/PMGifs/Genmoes/mitostat.html）。

200000bp。真核生物中还有其他一些类似叶绿体的细胞器。顶复门（Apicomplexa）的单细胞原生寄生动物中［如 Toxoplasma gondii（表 12-4）］具有更小一些的质体基因组。

表 12-4 叶绿体基因组选集

物种	通用名	编号	大小/bp
拟南芥(*Arabidopsis thaliana*)	拟南芥	NC_000932	154478
红藻(*Guillardia theta*)	红藻	NC_000926	121524
地钱，苔藓(*Marchantia polymorpha*)	地钱，苔藓	NC_001319	121024
烟草(*Nicotiana tabacum*)	烟草	NC_001879	155939
稻(*Oryza sativa*)	稻	NC_001320	134525
红藻(*Porphyra purpurea*)	红藻	NC_000925	191028
寄生虫(*Toxoplasma gondii*)	寄生虫(*Apicomplexan*)	NC_001799	34996
玉米(*Zea mays*)	玉米	NC_001666	140384

12.7.5 第一个真核染色体（1992 年）

第一个真核染色体的测序工作完成于 1992 年：啤酒酵母（*S. cerevisiae*）的Ⅲ号染色体（Oliver 等，1992）。序列包含了 182 个系统预测的开放式阅读框（对应于大于 100 个氨基酸的蛋白质），测序的 DNA 序列大小为 315kb。在已经识别的 182 个开放式阅读框中，只有 37 个和当时已知的基因相对应，另有 29 个显示与已知的基因存在着相似性。我们将在第 15 章中系统地研究这个基因组。

12.7.6 独立生存的生物体的全基因组（1995 年）

第一个完成测序的独立生存的生物体的基因组是流感细菌 *H. influenzae* Rd 的基因组（Fleischmann 等，1995）。它的大小是 1830138bp［即 1.8Mb］。这个生物体的基因组是在基因组研究所（TIGR：http://www.tigr.org）采用全基因组鸟枪（shotgun）测序法和装配策略（详见下文描述）完成测序的。

我们将这个细菌基因组描述为来自"独立生存的生物体"，是为了区别于来自病毒的基因组或者细胞器的基因组。病毒（参见第 13 章中相关描述）处于生命定义的边缘，而细胞器的基因组虽然通常认为来源于细菌，但是已经不再是独立的生命体了。

要在 NCBI 中研究这个基因组，首先访问 Entrez，点击“genome（基因组）”，然后点击左边侧栏中的“bacteria genome（细菌基因组）”，这时将列出超过 100 个细菌的全基因组；下拉到 *H. influenzae* 并点击它的编号［这时你还将看到其在 RefSeq 数据库中的编号（NC _ 000907），基因组大小（1830138bp），以及数据建立的时间（1995 年 7 月 25 日）］。这个生物体的页面包含大量关于基因和所编码的蛋白质的信息，另外还有蛋白质的功能分类预测（图 12-9）。分类方法、直系同源基因簇（clusters of orthologous genes，COG）将在第 14 章中作进一步讨论。起源关系（图 12-9，底部）显示这是一个位于 gamma 门中的细菌。与线粒体基因组（图 12-7）一样，*H. influenzae* 基因组的环形图表也是可以点击进入

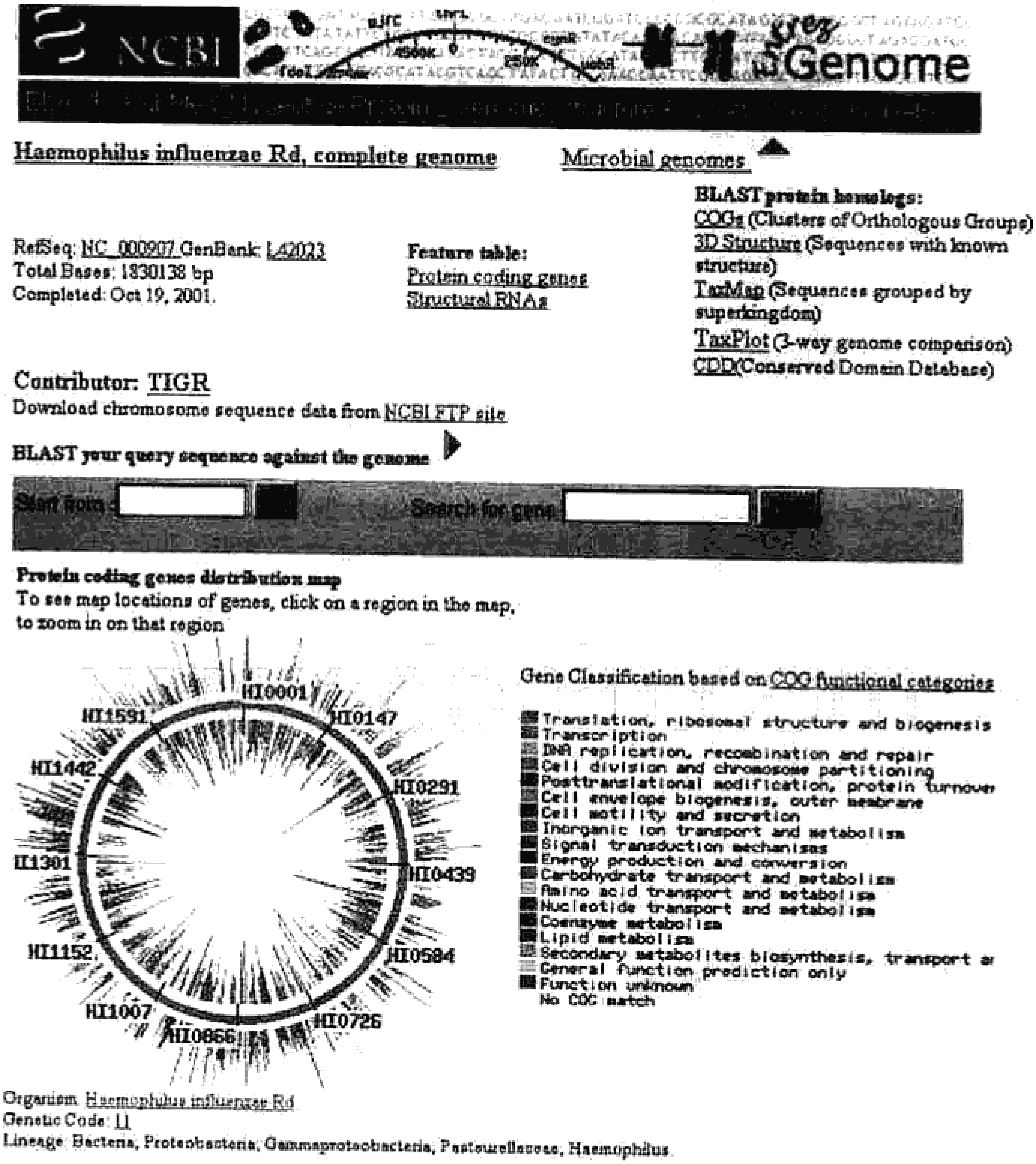


图 12-9　第一个完成基因组测序工作的独立生活的生物体 *H. influenzae* Rd 的 Entrez 基因组记录。这个记录是从 Entrez 基因组资源中点击其左侧栏的“bacteria（细菌）”得到的。记录的顶端包含了基因组的编号以及大小等信息。整个核苷酸序列都可以从这里下载。右上方是关于这个基因组所编码的 1709 个蛋白质的一些资源。从这里可以进行 *H. influenzae* 的 BLAST 搜索。记录的主要部分是由一个代表这个基因组的彩色圆环和一个基于 COG（见下文）的功能分类组成的，环形图表被点击以后（图 12-10）将显示这个基因组的基因和蛋白质的详细信息。底部是这个细菌的基因代码以及分类学起源，这个记录还包含了参考文献（图中没有显示出来），包括基因组研究所 Fleischmann 等在 1995 年的最初的基因组序列报告（见彩图）

的，从而得到其进一步的基因和蛋白质序列的详细研究数据（图 12-10）。

(a)

(b)

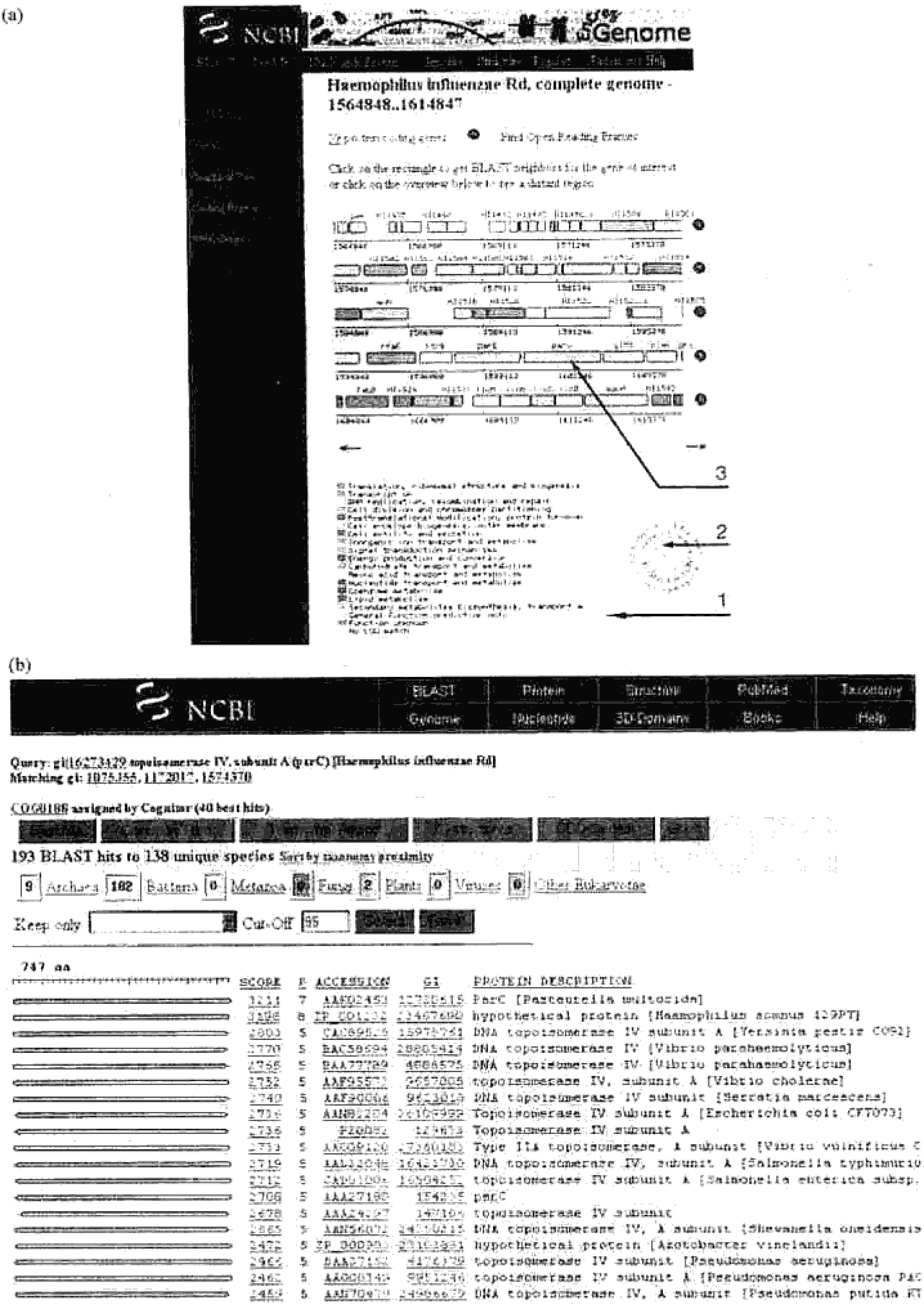


图 12-10 系统地研究全基因组的工具。(a) 点击 *H. influenzae* 基因组的环形图表（参见图 12-9），可以看见蛋白质编码基因的细节图。蛋白质会根据 COG 方案（图中箭头 1）用彩色来标识（详情参见表 8-9 以及第 14 章中的相关描述）。基因组的范围已经标出（图中箭头 2）。每个蛋白质都是可点击进入的。(b) 点击 parC（图中箭头 3）可以得到更多的信息，其中列出了一些可能的直系同源蛋白质序列的比对，包括几种细菌、病毒和植物的蛋白序列的比对

1995年底（表12-5），人们获得了第二个细菌基因组的完整的DNA序列，即 *Mycoplasma genitalium*（生殖道支原体）的DNA序列。值得注意的是，这是独立生活的生物体中已知的最小的基因组之一。我们将在第14章中详细研究这个细菌以及其他细菌。

表12-5 1995年已完成的基因组测序计划

生物体	大小/bp	编号	参考
Haemophilus influenzae Rd(流感细菌)	1830138	NC_000907	Fleischmann等,1995
Mycoplasma genitalium(生殖道支原体)	580074	NC_000908	Fraser等,1995

12.7.7 第一个真核基因组（1996年）

第一个真核生物（*S. cerevisiae*，一种啤酒酵母；参见第15章）(Goffeau等，1996）的完整的基因组在1996年完成测序（表12-6）。这个工作是在遍布欧洲、北美和日本的超过100个实验室600个研究人员的协作下共同完成的。如果想得到有关这个全基因组更进一步的信息，可以访问Entrez基因组网站，并通过在左边的侧栏中点击“prominent organisms（主要生物）”或者“eukaryota genome（真核基因组）”来获得。

表12-6 1996年已完成的基因组测序计划

生物体	大小/bp	编号	参考
甲烷球菌[*Methanococcus jannaschii*(A)]	1664976	NC_000909	Bult等,1996
肺炎霉浆菌[*Mycoplasma pneumoniae*(B)]	816394	NC_000912	Himmelreich等,1996;见Dandekar等,2000
集胞藻:*Synechocystis* PCC6803(B)	3573470	NC_000911	Kaneko等,1996
啤酒酵母:*Saccharomyces cerevisiae*(E)	12068	对每个染色体有不同编号	Goffeau等,1996

注：1. 大小不包括染色体外在部分（特别针对 *M. jannaschii*）。
2. A表示古细菌；B表示细菌；E表示真核生物。

12.7.8 更多的细菌和古细菌（1997年）

1997年，两个古细菌的全基因组序列报告产生（表12-7）。在5个被报告的细菌基因组

表12-7 1997年已完成的基因组测序计划

生物体	大小/bp	编号	参考
耐高温生物[*Archaeoglobus fulgidus*(A)]	2178400	NC_000917	Klenk等,1997
甲烷细菌[*Methanobacterium thermoautotrophicum*(A)]	1751377	NC_000916	Smith等,1997
枯草杆菌[*Bacillus subtilis*(B)]	4214814	NC_000964	Kunst等,1997
伯氏疏螺旋菌[*Borrelia burgdorferi*(B)]	910724	NC_001318	Fraser等,1997;Casjens等,2000
大肠杆菌[*Escherichia coli*(B)]	4639221	NC_000913	Blattner等,1997
幽门螺旋杆菌[*Helicobacter pylori*(B)]	1667867	NC_000915	Tomb等,1997;Alm等,1999

注：A表示古细菌；B表示细菌。

中，最著名的是大肠杆菌 *Escherichia coli*（Blattner 等，1997；Koonin 等，1997），它在将近十年的时间里都被视为是细菌学研究的生物体模型。它的基因组大小为 4.6Mb，编码超过 4200 个蛋白质，其中有 38%的功能在当时还没有得到识别。我们将在第 14 章详细研究这个细菌。

12.7.9 第一个多细胞生物的基因组（1998 年）

第一个完成基因组测序的多细胞生物是线虫 *Caenorhabditis elegans*——虽然从技术上说，测序仍未完全完成（因为存在很难解决的重复 DNA 的成分问题）。序列跨度为 97Mb，预测编码超过 19000 个基因（*C. elegans* 测序组织，1998）。

到 1998 年，已经完成测序工作的古细菌基因组又增加了 2 个，总共达到了 4 个（表 12-8）。另外还有 6 个细菌基因组也完成了测序。测序发现，在 20 世纪导致数千万人死于斑疹伤寒的 α 蛋白细菌 *Rickettsia prowazekii* 的基因组序列与真核生物线粒体的基因组十分接近（Andersson 等，1998）。

表 12-8 1998 年已完成的基因组测序计划

生物体	大小/bp	编号	参考
极地古细菌[*Pyrococcus horikoshii OT3*(A)]	1738505	NC_000961	Kawarabayasi 等，1998
超耐热菌[*Aquifex aeolicus*(B)]	1551335	NC_000918	Deckert 等，1998
沙眼衣原体[*Chlamydia trachomatis*(B)]	1042519	NC_000117	Stephens 等，1998
肺炎衣原体[*Chlamydophila pneumoniae*(B)]	1230230	NC_000922	
结核杆菌[*Mycobacterium tuberculosis*(B)]	4411529	NC_000962	Cole 等，1998
普氏立克次体[*Rickettsia prowazekii*(B)]	1111523	NC_000963	Andersson 等，1998
梅毒螺旋体[*Treponema pallidum*(B)]	1138011	NC_000919	Fraser 等，1998
线虫[*Caenorhabditis elegans*(E)]	97000000	AE000001 AE000002 AE000003 AE000004 AE000005 AE000006	The C. elegans Sequencing Consortium(C. elegans 测序组织)，1998

注：A 表示古细菌；B 表示细菌；E 表示真核生物。

12.7.10 人类染色体（1999 年）

1999 年，人类第 22 号染色体常染色质部分的序列发表了（表 12-9）(Dunham 等，1999)。这是第一个被完全测序的人类染色体。我们将在第 17 章详细叙述人类基因组的相关内容。

表 12-9 1999 年已完成的基因组测序计划

生物体	大小/bp	编号	参考
嗜热古生菌[*Aeropyrum pernix*(A)]	1669695	NC_000854	Kawarabayasi 等，1999
Thermoplasma volvanium GSS1(A)	1584804	NC_002689	Kawashima 等，1999
肺炎衣原体[*Chlamydia pneumoniae*(B)]	1229858	NC_002179	Kalman 等，1999

续表

生物体	大小/bp	编号	参考
耐辐射奇异球菌[*Deinococcus radiodurans*(B)]	2648638	NC_001263	White 等，1999
	412348	NC_001264	White 等，1999
幽门螺旋杆菌[*Helicobacter pylori* J99(B)]	1643831	NC_000921	Alm 等，1999；Tomb 等，1997
Thermotoga maritima (B)	1860725	NC_000853	Nelson 等，1999
人类[*Homo sapiens* 第 22 号染色体(E)]	33792315	NC_001454，10 个其他 contig	Dunham 等，1999

注：A 表示古细菌；B 表示细菌；E 表示真核生物。

12.7.11 蝇、植物和人类 21 号染色体（2000 年）

2000 年，果蝇 *Drosophila melanogaster* 和植物 *A. thaliana* 均被报告基因组测序已经完成，使得真核生物基因组的数量增加到四个（还有一个酵母和一个蠕虫）（表 12-10）。果蝇 *Drosophila* 的序列是由 Celera 基因公司（Celera Genomics）以及伯克利果蝇基因组计划（Berkeley Drosophila Genome Project，BDGP）的科学家们获得的（Adams 等，2000）。大约有 13500 个基因已被注解。*Arabidopsis* 是芥科植物拟南芥中的一种。它紧凑简洁的基因组使其成为了植物基因组的模式生物。

我们将在第 16 章中详细叙述这些真核生物以及其他真核生物基因组的相关研究成果。

同样在 2000 年，人类第 21 号染色体成为了第二个被报告的人类染色体序列（Hattori 等，2000）。这是人类常染色体中最小的一条。这条染色体的额外复制将导致唐氏综合征，是一种最为常见的遗传性智力低下病症。

与此同时，细菌基因组的测序工作仍在继续，出现了许多令人惊奇的特性。细菌性脑膜炎的致病菌——脑脊髓膜炎双球菌（*Neisseria meningitidis*）的基因组包含上百个重复的成分。这种重复现象在真核生物中比较典型。铜绿假单胞菌（*Pseudomonas aeruginosa*）的基因组大小为 6.3Mb，是当时已测序的最大的细菌基因组（Stover 等，2000）。

在古细菌中，嗜酸热支原体（*Thermoplasma acidophilum*）的基因组测序完毕（Ruepp 等，2001）。这种生物在 59℃ pH2 的环境中最为活跃。引人注意的是，它曾经从硫磺矿硫化叶菌（*Sulfolobus solfataricus*）那里得到了大量侧向转移的基因，而后者是一种从系统发生学上与前者相距较远的一种古细菌，两者仅仅是一同生活在煤堆这种相同的生态环境中而已。

我们将在第 14 章中详细讨论水平基因转移的相关问题。

12.7.12 人类基因组序列的草图（2001 年）

有两个小组完成并发表了人类基因组的草图——即由国际人类基因组测序组织（International Human Genome Sequencing Consortium，IHGS，2001 年）以及 Celera 基因公司领导的一个组织完成（表 12-11）（Venter 等，2001）。这两份报告都得出相同的结论，即基因组中大约包含了 30000～40000 个蛋白质编码基因——一个没有被人们预先料到的小数目。对人类基因组序列的分析将对人类生物学的各个方面的发展都产生广泛的影响（参见第 17 章）。

表 12-10 2000年已完成的基因组测序计划

生物体	大小/bp	编号	参考
极端嗜盐菌[*Halobacterium* sp. NRC-1(A)]	2014239	NC_002607	Ng 等,2000
嗜酸热支原体[*Thermoplasma acidophilum*(A)]	1564906	NC_002578	Ruepp 等,2000
Bacillus halodurans(B)	4202353	BA000004	Takami 等,2000
蚜虫内共生菌[*Buchnera* sp. APS(B)]	640681	NC_002528	
豚肠弯曲杆菌[*Campylobacter jejuni*(B)]	1641481	NC_002163	Parkhill 等,2000
Chlamydia muridarum(B)	1069412	NC_002178	Read 等,2000
	1228267	NC_002491	
肺炎衣原体[*Chlamydia pneumoniae* AR39(B)]	1229853	NC_002179	Read 等,2000
沙眼衣原体[*Chlamydia trachomatis*(B)]	1069412	NC_000117	Read 等,2000;见 Stephens 等,1998
脑脊髓膜炎双球菌[*Neisseria meningitidis* MC58(B)]	2272351	NC_002183	Tettelin 等,2000
脑脊髓膜炎双球菌[*Neisseria meningitidis* Z2491(B)]	2184406	NC_002203	Parkhill 等,2000
铜绿假单胞菌[*Pseudomonas aeruginosa*(B)]	6264403	NC_002516	Stover 等,2000
解脲支原体[*Ureaplasma urealyticum*(B)]	751719	NC_002162	Glass 等,2000
霍乱弧菌[*Vibrio cholerae*(B)]	2961149	NC_002505	Heidelberg 等,2000
	1072315	NC_002506	Heidelberg 等,2000
苛养木杆菌[*Xylella fastidiosa*(B)]	2679306	NC_002488	Simpson 等,2000
果蝇[*Drosophila melanogaster*(E)]	137000000	NC_004354 NT_003779 NT_003778 NT_037436 NT_033777 NC_004353	Adams 等,2000
拟南芥[*Arabidopsis thaliana*(E)]	125000000	NC_003070 NC_003071 NC_003074 NC_003075 NC_003076	拟南芥属基因组计划(The Arabidopsis Genome Initiative),2000
人类[*Homo sapiens* 21 号染色体(E)]	33800000	NT_002836 NT_002835 NT_003545 NT_001715 NT_001035	Hattori 等,2000

注:A 表示古细菌;B 表示细菌;E 表示真核生物。

表 12-11 2001年已完成的基因组测序计划

生物体	大小/bp	编号	参考
一种生活在深海热管道中的生物[*Pyrococcus abyssi*(A)]	1765118	NC_000868	R. Heilig,2001
硫叶菌[*Sulfolobus solfataricus*(A)]	2992245	NC_002754	She 等,2001
Sulfolobus tokodaii(A)	2694765	NC_003106	Kawarabayasi 等,2001
新月柄杆菌[*Caulobacter crescentus*(B)]	4016942	NC_002696	Nierman 等,2001
大肠杆菌[*Escherichia coli* 0157:H7(B)]	5498450	NC_002695	Perna 等,2001;Hayashi 等,2001;见 Blattner 等,1997
	5528970	AE005174	Perna 等,2001
麻风分枝杆菌[*Mycobacterium leprae*(B)]	3268203	NC_002677	Cole 等,2001
肺支原体[*Mycoplasma pulmonis*(B)]	963879	NC_002771	Chambaude 等,2001
出血性巴斯德杆菌[*Pasterurella multocida*(B)]	2257487	AE004439	May 等,2001
苜蓿中华根瘤菌[*Sinorhizobium meliloti*(B)]	6.7×10^6	NC_003047	Galibert 等,2001
肺炎链球菌[*Streptococcus pneumoniae*(B)]	2160837	AE005672	Tettelin 等,2001
化脓性链球菌[*Streptococcus pyogenes*(B)]	1852442	AE004092	Ferretti 等,2001
家兔脑内原虫病[*Encephalitozoon cuniculi*(E)]	2.5×10^6	AL391737 和 AL590442 到 AL590451	Katinka 等,2001
隐滴虫类[*Guillardia theta* 类核体基因组(E)]	551264	NC_002751	Douglas 等,2001
人类[*Homo sapiens*(E)]	3.3×10^9	多个编号	IHGSC,2001;Venter 等,2001

注:A 表示古细菌;B 表示细菌;E 表示真核生物。

新测序完毕的细菌基因组继续给我们带来了有意义的特征。肺炎霉浆菌（*Mycoplasma pulmonis*）是拥有最小的鸟嘌呤-胞嘧啶（GC）含量之一的生物，其GC含量为26.6%（Chambaud等，2001）。麻风病的致病菌——麻风分枝杆菌（*Mycobacterium leprae*）的基因组曾经发生了大量的基因失活，目前只有一半的基因组仍然是编码基因（Cole等，2001）。对出血性巴斯德杆菌（*Pasterurella multocida*）的基因组进行分析得出，蛋白细菌包括流感细菌（*H. influenzae*）和大肠杆菌（*E. coli*）以及其他致病的革兰阴性细菌的γ亚门的扩张出现在大约6.8亿年前（May等，2001）。苜蓿中华根瘤菌（*Sinorhizobium meliloti*）的基因组包含一个环形染色体以及两个附加的巨型质体（Galibert等，2001）。这三项加起来总共6.7Mb，令我们对细菌基因组结构的多样性有了新的认识。

隐藻是一种将一个独特的真核细胞（带细胞核的红藻）包含于另一个细胞内的藻类（框16-2）。这种独特的结构源自于古代两个生物体的一次进化融合。红藻的细胞核被称为类核体，是已知的真核基因组中基因最为密集的细胞核。测序后（Douglas等，2001）发现它的基因组密布（每977bp一个基因）了许多极短的非编码区域。

12.7.13 基因组测序的不断完成（2002年）

2002年，又有数十个微生物基因组测序完毕。在真核生物（表12-12）中，发现了裂殖酵母*Schizosaccharomyces pombe*具有最小的蛋白质编码基因数目（4824）（Wood等，2002）。关于疟原虫*Plasmodium falciparum*以及它的寄主蚊子*Anopheles gambiae*的基因组的报告均已产生（Holt等，2002）。另外，啮齿动物疟原虫*Plasmodium yoelii*的基因组已被测定并与*P. falciparum*作了对比（Carlton等，2002）。这些计划将在第16章中详细叙述。

表12-12 2002年已完成的真核生物基因组测序计划

生物体	大小/Mb	编号	参考
蚊子[*Anopheles gambiae*(E)]	278	AAAB00000000	Holt等,2002
疟原虫[*Plasmodium falciparum*(E)]	22.8	NC_002375	Gardner等,2002
Plasmodium yoelii yoelii(E)	23.1	AABL00000000	Carlton等,2002
裂殖酵母[*Schizosaccharomyces pombe*(E)]	13.8	NC_003424 NC_003423 NC_003421	Wood等,2002

注：E表示真核生物。

12.8 基因组分析总览

我们已经按时间顺序考察了已完成的基因组计划。还有很多与基因组测序有关的问题。例如，哪个基因组已被测序了？它是怎么完成的？这些基因组有多大？基因组的DNA测序完毕，它的主要特征是什么（如基因、调控区域、重复成分），以及它们是如何测定的？甚至当我们正在学习问何种问题以及处理这些问题可以使用何种工具时，序列分析的目标本身也在发展。

接下来将把基因组序列分析过程划分为几个相对独立的阶段（从选择适合用于DNA测序的基因组到基因组的注解），并一一加以研究（图12-11）。斯坦（Stein，2001）和库宁（Koonin，2001）曾就计算基因组学写过很多相关的评论。

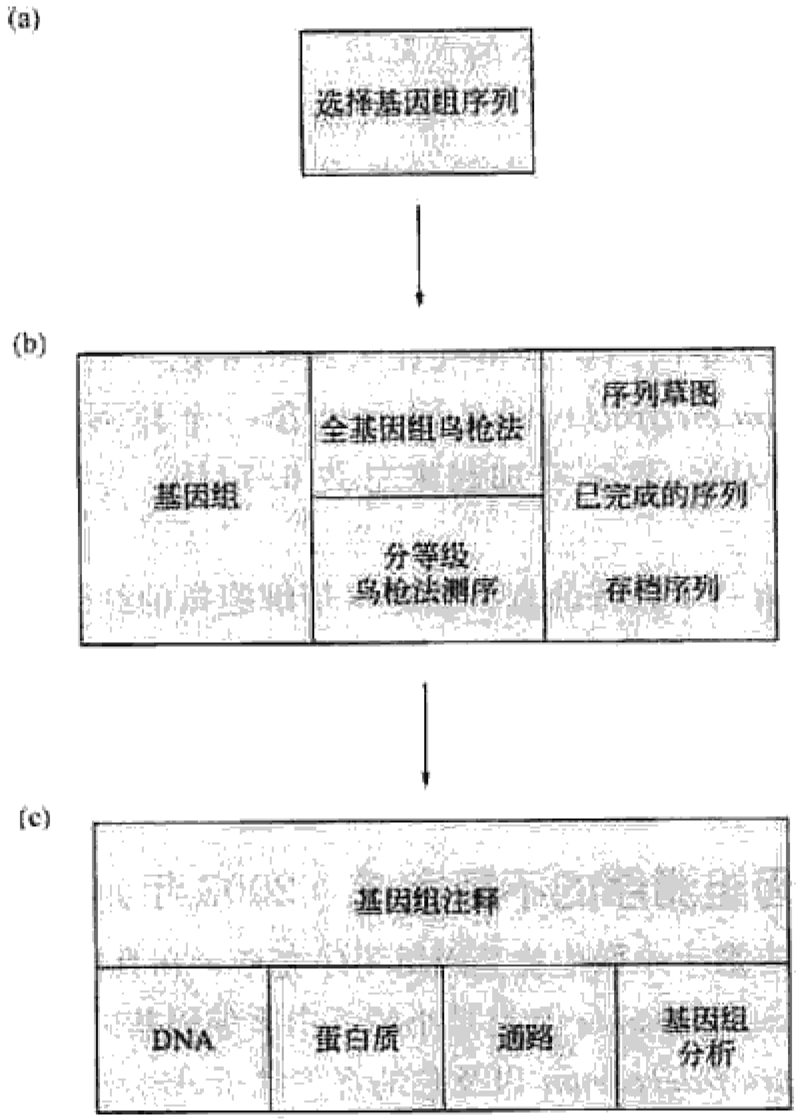


图 12-11 基因组测序和分析过程总览。(a) 要进行测序的基因组涉及到花费的代价，生物和疾病方面的意义等因素。(b) 基因组测序中心可以采用多种手段来进行基因组测序，比如全基因组鸟枪测序法、分级全基因组鸟枪测序法，以及两者同时采用。DNA测序分阶段进行，包括生成草图完成序列、结果存档。(c) 完成后的序列将在DNA的层面上（比如标识出重复的成分、核苷酸的构成以及蛋白质编码基因）、在预测的蛋白质层面上以及在预测的细胞通路层面上加以注解。另外，还可以进行一系列基因组层面上的分析，如进行基因组或系统发生学比较（见下文的详细描述）

12.8.1 选择用于测序的基因组

选择哪个基因组进行测序取决于几方面的因素。随着时间的推移，选取的准则也会发生改变，因为技术的进步可以使成本消耗降低，并且基因组测序中心会在这些努力下不断积累实践经验。

12.8.1.1 基因组的大小

对于微生物基因组来说，典型的大小通常为数百万个碱基对，单个测序中心就已经拥有完成整个计划所需的全部资源。而对于更大的（尤其是真核生物）基因组，通常需要建立国际间合作以分担工作（详情参见第16章）。对于人类基因组计划（参见第17章），在过去的几年中，按一天24h算，大约每秒钟就完成1000个碱基对的测序。

到2003年6月，卫康信托桑格研究所（Wellcome Trust Sanger Institute，网址：http://www.wellcome.ac.uk/）已经资助了包括数十种生物的超过2Gb的DNA测序工作（http://www.sanger.ac.uk/Info/Statistics/）。

12.8.1.2 消耗的代价

通过一个公共组织来完成人类基因组序列草图的世界范围的消耗大约为3亿美元。支持基因组测序的主要资金来源是美国国家卫生研究院（National Institutes of Health，Bethesda，Maryland）和卫康信托（Wellcome Trust，England）两家机构。

12.8.1.3 有关人类疾病

例如，通过对黑猩猩基因组的测序，我们可以发现为什么这些动物不会感染某些能够感染人类的疾病，例如疟疾和艾滋病（参见第18章）。

12.8.1.4 有关基础生物的问题

例如，鸡提供了一个非哺乳类的脊椎系统研究模型，广泛用于对发育的研究中。而对原生动物基因组的分析可以揭示真核生物的进化过程。

12.8.1.5 有关农业

普遍认为鸡、奶牛和蜜蜂等的基因组序列将在许多方面有益于农业的发展，如研究如何保护这些生物以免受疾病侵害的方法等。

美国国家人类基因组研究所（National Human Genome Research Institute，NHGRI）是赞助基因组测序计划的联邦机构之一。NHGRI 提出议案（称为“白皮书”）供各个研究小组申请资金以及完成基因组测序计划，并提供选择某个特定生物体的基本原理（表 12-13）。可以从 NHGRI 的相关网站中读到这些议案。

表 12-13 基因组测序提案（2002 年）

生物体	物种	优先级
鸡	原鸡(*Gallus gallus*)	高
黑猩猩	黑猩猩(*Pan troglodytes*)	高
奶牛	温带牛(*Bos taurus*)	高
狗	犬(*Canis familiaris*)	高
真菌	共 15 个物种	高
蜜蜂	西方蜜蜂(*Apis mellifera*)	高
原生动物纤毛虫(*Oxytricha*)	一种有纤毛的原生动物(*Oxytricha trifallax*)	高
海胆	紫海胆(*Strongylocentrotus purpuratus*)	高
单细胞纤毛虫(*Tetrahymena*)	四膜虫(*Tetrahymena thermophila*)	高
低级多细胞动物(*Trichoplax*)	低等多细胞动物(*Trichoplax adhaerens*)	中
恒河猴	猕猴(*Macaca mulatta*)	中

注：改编自 NIH 的美国国家人类基因组研究所网站（http://www.genome.gov/page.cfm?pageID=10002154），这个网站提供了研究小组的每个测序计划的“白皮书”议案的链接。

12.8.2 应当对物种的一个、几个还是多个个体测序

从根本上说，为了将基因型和表型联系在一起，从物种的多个个体中测定整个基因组序列是十分重要的。以人类的情况为例（参见第 17 章），公共组织的人类基因组计划从一开始就涉及到对各种种族背景的多个个体（包括男性、女性）的基因组 DNA 的测序和分析。

对于诸如人类免疫缺陷病毒（HIV-1 和 HIV-2）这样的病毒，由于病毒的 DNA 总是在非常迅速地不断变化，因而必须对上千个独立的个体进行测序（参见第 13 章）。病毒基因组极小（<10kb），因此这种做法是可行的。在某些情况下，比较不同的细菌菌株可以发现为什么一个是无害的而另一个却是高度致病的（表 14-10）。这种比较曾在大肠杆菌（*E. coli*）的一种普遍存活于人类消化道中的菌株以及另一种导致严重甚至致命疾病的菌株中进行（参见第 14 章）。

12.8.3 基因组究竟有多大

图 12-12 显示了各个基因组大小的比较情况。病毒基因组的范围为 1～350kb（参见第 13 章）。在单倍体基因组（如细菌基因组，参见第 14 章）中，基因组的大小（或 C 值）就是基因组中 DNA 的总量。在双倍体或者多倍体生物中，基因组的大小就是无重复的单倍体

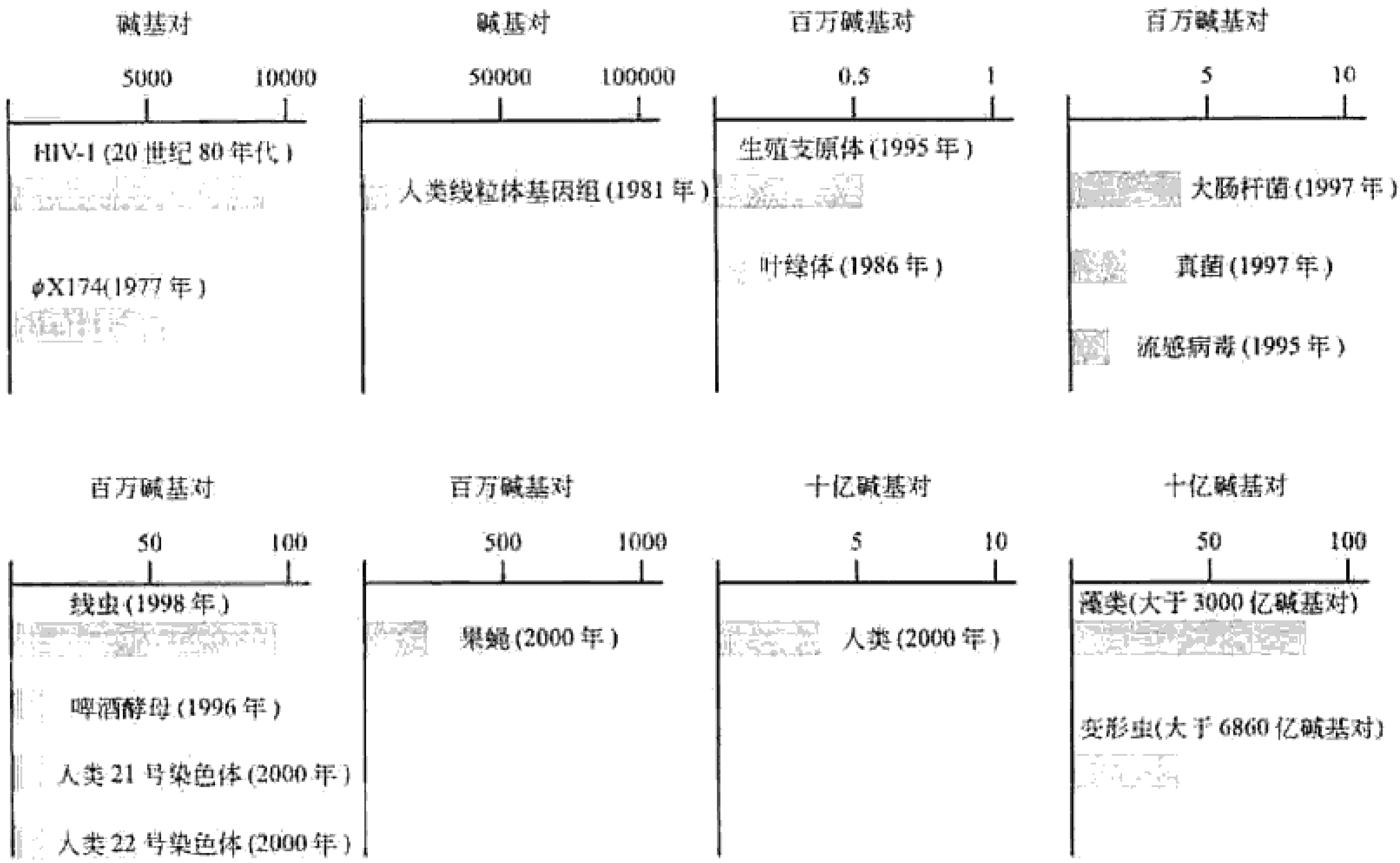


图 12-12 多个不同基因组的大小比较。每幅图的 x 轴代表了一个10倍的尺度变化。对于细菌，基因组大小的范围从仅仅580000bp（已测序的最小的基因组是生殖器支原体 *M. genitalium* 的基因组，含470个蛋白质编码基因）到藻青菌基因组的13Mb，跨越了22倍的变化范围。对于真核基因组，范围从一些真菌的8Mb到一些变形虫的686Gb。这个范围上下限相差超过了75000倍，并被称为 *C* 值矛盾（*C* value paradox，参见第16章）。*C* 值是指基因组中DNA的总含量，而 *C* 值矛盾则指真核生物的复杂度与其基因组DNA总量的关系

基因组（比如精子细胞核基因组）的DNA总量。细菌基因组的大小在一个上下浮动大约22倍的范围内变化，从大约500000bp（生殖器支原体 *M. genitalium*）到13Mb（目前已测序的最大的原核生物基因组 *Streptomyces coelicolor* 的大小是8.7Mb）。

在真核生物中，基因组的大小可以相差大约75000倍，即从一些真菌的8Mb到一些变形虫的686Gb。所谓的 *C* 值矛盾（*C* value paradox）指的是有些生物虽然 *C* 值很大，但是形态上却相当简单，蛋白质编码基因的数量也处于较一般的水平。我们将在第16章详细讨论 *C* 值矛盾。

12.8.4 基因组测序中心

大尺度的测序是在世界范围的各个中心共同操控下完成的。有20个测序中心为人类基因组草图的完成做出了贡献（2001年）（表17-6）。这些中心也受到了NIH和EBI的支持。所有这些中心同时也参与了其他生物体基因组的测序工作。

除了表17-6所列举的中心以外，还有许多其他重要的基因组测序中心。它们包括曾经贡献了数十个原核和真核基因组测序计划的基因组研究所（The Institute for Genomics Research，TIGR）以及在2001年曾生成人类基因组草图的Celera基因公司（Celera Genomics）。

基因组测序中心的列表可以从 NCBI 的网站（http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/links.html）、人类基因组图谱计划资源中心（UK Human Genome Mapping Project Resource Centre，http://www.hgmp.mrc.ac.uk/GenomeWeb/genome-centres.html）以及 http://igweb.integratedgenomics.com/GOLD/Genomiclinks.html 上得到。

TIGR 的网站在 http://www.tigr.org/，Celera 的在 http://www.celera.com/。两者都位于马里兰的 Rockville。

12.8.5 基因组测序：策略

基因组测序目前主要有两种方法。第一种是全基因组鸟枪测序法（whole-genome shotgun，WGS）。桑格首先采用了这个方法对抗菌素 ϕX174 进行了测序：从基因组 DNA 中随机地选取一些片段，然后互相隔离并进行测序，最后装配在一起形成一个完整的序列。而最先将这个方法应用于整个生物体基因组的是约翰·霍普金斯（Johns Hopkins）的汉密尔顿·史密斯（Hamilton O. Smith）和 TIGR 的克雷格·文特尔（J. Craig Venter，后来去了 Celera 基因公司）。

全基因组鸟枪方法已经成功应用于大多数小基因组中（例如病毒、细菌和古细菌，以及没有大量重复 DNA 的真核生物的基因组）。首先将基因组 DNA 从生物体上分离并进行机械剪切（或者用限制酶来分解）。然后将片段亚克隆至小嵌件库（如 2kb 的片段）和大嵌件库中（如 10～20kb）。从两头开始对克隆进行测序（即正义链和负义链），然后测序结果被装配起来。一次典型的测序反应大约产生 500 个碱基对的序列数据。这些小量的序列装配在一起就产生了连续的转录本（“contigs，叠连群”），最终形成了完整的基因组。这个策略被大量应用于表 12-5～表 12-12 中描述的细菌和古细菌的测序计划中。流感细菌 *H. influenzae* 的基因组在 1995 年用全基因组鸟枪方法完成了测序（表 12-5）。

有关基因组测序的术语见表 12-14。

第二种相关的方法是“分级鸟枪测序法”（图 12-13）。基因组 DNA 被分解并亚克隆至 BAC 库中。这些库包含了大的嵌件（100～500kb）。另外，也可以生成小一些的黏粒库（嵌件大小约为 50kb）或者质粒库（2～10kb）。跟 WGS 测序法不同的是，这种分级策略所采用的克隆（叠连群）最终被映射到已知的染色体位置上。这样，序列的装配只需关注一个小范围内的基因组即可。这种方法已被应用于大部分大的真核基因组中，包括公共组织版本的人类基因组计划。

异染色质中的区域包含大量高度重复的 DNA 片段（参见第 16 章），无法用 WGS 或者分级方法进行有效的测序。Skaletsky 等（2003 年）使用了另外一种迭代映射和测序技术测定了人类 Y 染色体中的重复序列。

不管叠连群源自哪个染色体区域，WGS 方法都需要进行复杂的计算才能将它们组合到一起。有人认为这种方法不可能用于大的真核基因组。然而，这个方法结合分级方法已经应用于 120Mb 的果蝇（*D. melanogaster*）基因组（Adams 等，2000）以及人类基因组中（Weber 和 Myers 等，1997；Venter 等，2001）(参见第 17 章)。

12.8.6 基因组序列数据：从未完成到完成

Phred 和 Phrap（见下文）可以从 http://www.phrap.org/获取。它们运行在 UNIX 类的系统中。

原始的 DNA 数据堆放在 NCBI 和 EBI 的 Trace Archives 等数据库中（参见下文）。这些数据主要是供计算机软件读取的，如

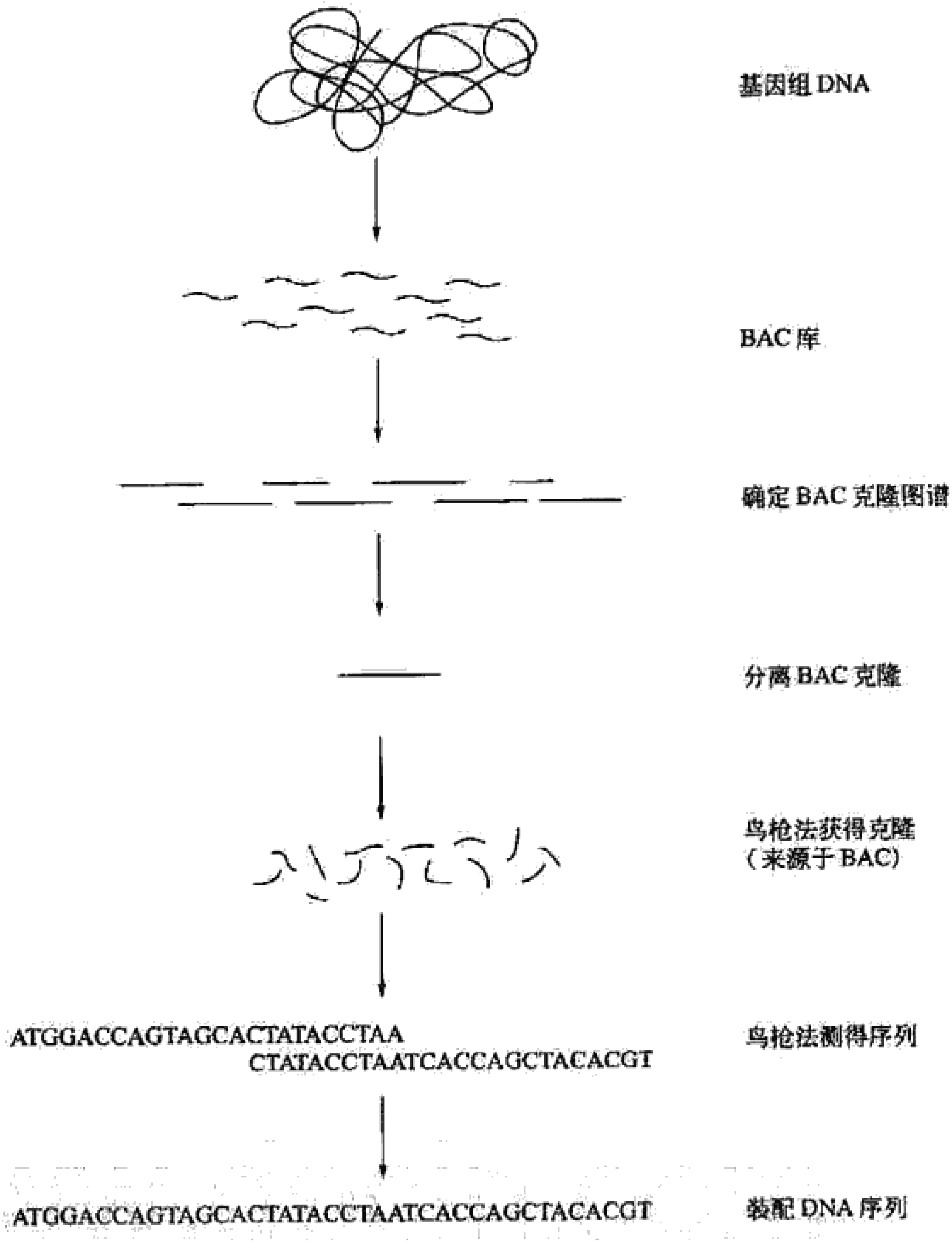


图 12-13 分级鸟枪测序策略的示意。基因组 DNA 被从目标生物体中分离、分解成片段并插入到 BAC 库中。每个 BAC 克隆大约为 100～500kb。BAC 已按顺序排列（映射）。各个 BAC 克隆被分解为更小的 cDNA 克隆并测序。各个测序反应的典型值为 300～700 个核苷酸。这些"鸟枪序列"被装配在一起。对这个过程将在图 17-14 中做更进一步的详细讲述

修改自 IHGSC（2001 年，第 863 页）。数据经授权使用

Phred 程序。这个程序解析 DNA 测序机所测序的碱基并进一步评估数据的质量。然后 Phred 以某种格式［如 FASTA 格式（图 2-10）］将序列记录下来，留作进一步分析。

在 NCBI 中，原始的基因组 DNA 数据可以通过高通量基因组（high-throughput genomic，HTG）序列分类获取。每个数据项都有固定编号。高通量基因组数据库包含四个阶段的序列数据（表 12-15）。阶段 0 的数据一般是源自单个黏粒或 BAC 的数据。它们通常包含测序错误以及大小不确定的间隙。尽管如此，这种形式的数据对于科学界仍然有极大的用处。例如，如果正在进行 BLAST 搜索，试图寻找查询对象的新的同源物，高通量基因组（HTG）分类就可能包含有用的信息。阶段 1 可以包含一个大的克隆（如一个 BAC 克隆）的多个叠连群的序列。其中，叠连群的顺序以及它们的方向（正义链或负义链）均未知。序列被标记为未完成，并且仍含有间隙。

在完成状态中（阶段 2），叠连群已被正确地定序和定向，错误率必须在 10^{-4} 或以下。

表 12-14 基因组测序计划使用的术语

术　语	定　义
BAC 终端序列(BAC end sequence)	细菌人工染色体(bacterial artificial chromosome,BAC)的终端序列已被测序并提交至 GenBank;BAC 的内部序列可能无法得到。只要得到了同一个 BAC 两端的终端序列就可以利用它对叠连群进行定序并形成骨架
叠连群(contig)	可以形成一条完整序列的一系列重叠的克隆或序列。NCBI 的叠连群记录代表从多个克隆序列中构建的连续序列。这些记录可能包含草图或者完成序列,也可能包含序列间隙(在单个克隆中)或者跨越了别的未测序克隆的多个克隆之间的间隙
草图序列(draft sequence)	在人类基因组测序计划中的定义:在鸟枪测序阶段的 Phred Q20 中,至少覆盖预计克隆嵌件的 3~4 层。注意"草图"的确切定义在别的基因组计划中会有不同。克隆序列可能包含以间隙分割的多块序列。这些小块的真实顺序和方向可能仍然未知
完成序列(finished sequence)	克隆嵌件形成连续的高质量的错误率小于 0.01%的编码链。序列中通常不包含间隙
片段(fragment)	克隆序列中的一段连续延伸的序列,中间不包含间隙、载体或者其他污染序列
合并(meld)	当两个或多个重叠的序列占据同一个比对区域时,将这些序列合并在一起组成一个长的序列
定序和定向(order and orientation)	利用序列的重叠信息对一个大的克隆序列中的多个片段重新定序并定向(order and orient,ONO)
骨架(scaffold)	在染色体中排序以后的叠连群

注:改编自 http://www.ncbi.nlm.nih.gov/genome/guide/build.html。

表 12-15 GenBank 的高通量基因组记录中定义的四个阶段

状　态	位　置	定　义
阶段 0	高通量基因组分类	单个克隆(非叠连群)的单次或少次读取
阶段 1	高通量基因组分类	未完成,仍未定序、定向的叠连群,含间隙
阶段 2	高通量基因组分类	未完成,已定序、定向的叠连群,含或不含间隙
阶段 3	主分类	完成,不含间隙(有或没有注解)

注:来自 http://www.ncbi.nlm.nih.gov/HTGS/。

> GenBank 中的阶段 1、阶段 2 和阶段 3 的序列示例,见 http://www.ncbi.nlm.nih.gov/HTGS/examples.html。

装配过程牵涉到各个序列(阶段 0)的收集、间隙的填充以及错误率的降低。这个过程可以通过采用多种软件包来实现,如 Phrap(以及它的图形查看器 Consed)、Assembler 和 Sequencher。无论是全基因组测序法或者分级法,在鸟枪阶段完成以后,下一步都要将叠连群装配在一起。这一步是通过一个称为完成的过程来实现的。完成的目标就是要识别小块之间的间隙并填充。在理想的情况下,这个过程最终将产生一个跨越所有叠连群的单条连续的 DNA 序列。完成过程可以由专家手工进行,也可以通过程序(如 Autofinish)以自动的方式进行(Gordon 等,2001)。

基因组测序计划常常深深依赖表达序列标签(expressed sequenced tags,ESTs)以帮助蛋白质编码基因的界定。被表达的转录本(即 RNA 分子)先被转化为 cDNA,然后加入库中并被测序。这些 cDNA 就是表达序列标签(ESTs)。如果它们不揭示有关对应的基因组 DNA 的一些信息(如内含子序列),则对识别被表达的基因就没有什么价值(图 12-17)。

12.8.7 一个基因组什么时候测序完毕?

一般地,为了使基因组尽可能完全地被测序,基因组应当进行 5~10 层覆盖的测序。技术上面临的最大挑战是如何处理真核基因组的重复 DNA 的长片区域中的序列(见下文描

述）。目前，人类基因组是惟一的对大范围重复DNA序列进行仔细测序的基因组。

将已测序的DNA的估算量写成覆盖层数（fold coverage）的函数（表12-16）。兰德和沃特曼（Lander和Waterman，1988）给出了一个碱基没有被测序的概率为

$$P_0 = e^{-c} \quad (12\text{-}1)$$

式中，c是覆盖层数，可以写为

$$c = \frac{LN}{G} \quad (12\text{-}2)$$

式中，LN是测序的碱基总数；L是每次读取的长度；N是读取的次数；e是常数2.718。从以上公式可以得出，要达到万分之一（0.01%）的错误率，理论上说需要9层基因组覆盖。如果是5层覆盖，错误率大约为0.6%。

表12-16 根据方程（12-1）得出的一个碱基被测序的概率

覆盖层数	P_0	没有被测序的概率/%	已被测序的概率/%
0.25	$e^{-0.25}$	78	22
0.5	$e^{-0.5}$	61	39
0.75	$e^{-0.75}$	47	53
1	e^{-1}	37	63
2	e^{-2}	13.5	87.5
3	e^{-3}	5	95
4	e^{-4}	1.8	98.2
5	e^{-5}	0.6	99.4
6	e^{-6}	0.25	99.75
7	e^{-7}	0.09	99.91
8	e^{-8}	0.03	99.97
9	e^{-9}	0.01	99.99
10	e^{-10}	0.005	99.995

注：改编自 http://www.genome.ou.edu/poisson.calc.html 以及 Lander 和 Waterman（1998）。

某些基因组的DNA区域可能产生不确定的测序结果。这种现象可能出现在GC含量极高或极低或者二级结构很特别的区域中。这些区域通常通过使用特定的低核苷酸引物对问题区域直接进行测序来解决（"完成"）(Tettelin等，1999)。疟原虫*P. falciparum*的测序工作尤为困难，因为其基因组的80%都是由腺嘌呤（A）和胸腺嘧啶（T）组成的（Gardner等，2002）。因此，要对这个23Mb的基因组进行测序需要一种基于染色体的方法（参见第16章）。

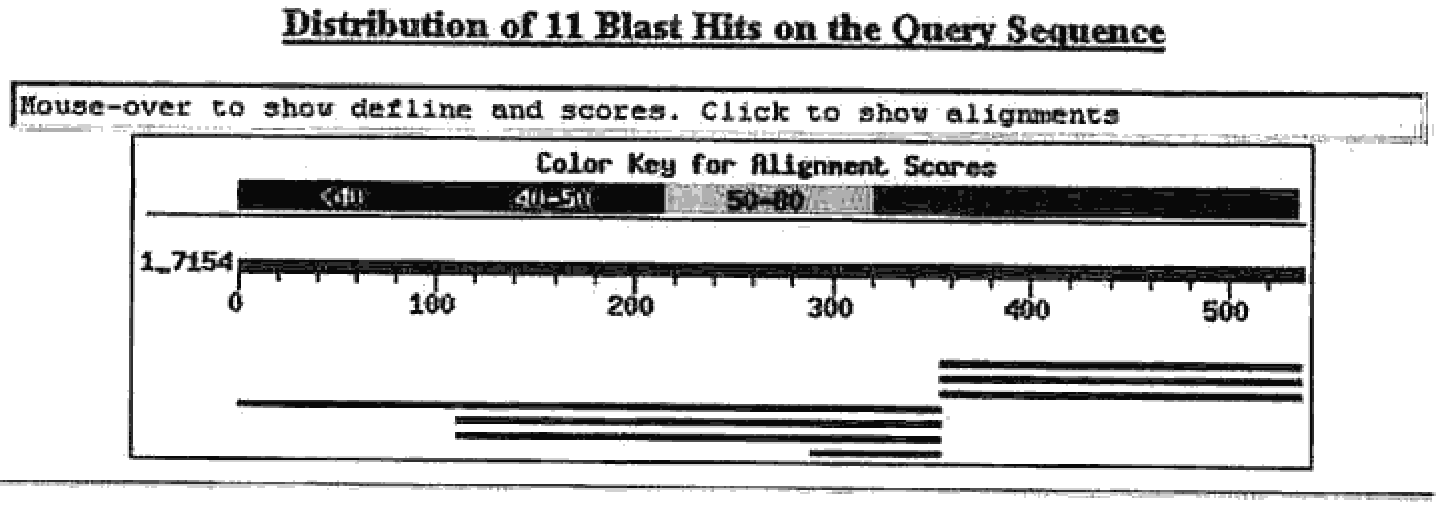
12.8.8 基因组序列的数据仓库

一个专门的Trace档案BLAST服务器可以从这里访问：http://www.ncbi.nlm.nih.gov/blast/mmtrace.html。

多种生物的基因测序计划的原始序列数据已经堆积在NCBI和EBI的Trace档案（Trace Archive）中。档案中的所有条目都有一个Trace识别号（Trace Identifier，Ti）。档案可以通过多种标准搜索（如通过Ti或测序中心或BLAST来查询）。

例如以小鼠的RBP克隆（M74527）搜索小鼠的Trace档案，输出结果包含了多个Ti匹配［图12-14(a)］。点击某个Ti记录的链接，可以得到FASTA格式文件的或者用以跟踪测序DNA时的染料终止反应的序列

(a)



Sequences producing significant alignments:		Score (bits)	E Value
gnl|ti|328428	G10P603056FB11.T0	329	5e-87
gnl|ti|9569624	G10P611529FH1.T0	315	6e-83
gnl|ti|20580100	jst20g07.g1	274	1e-70
gnl|ti|37502429	G10P648090RG8.T0	247	3e-62
gnl|ti|16878054	jip88d10.b1	247	3e-62
gnl|ti|9499554	G10P610230FE10.T0	247	3e-62
gnl|ti|18312529	G10P634600RC5.T0	120	5e-24

(b)

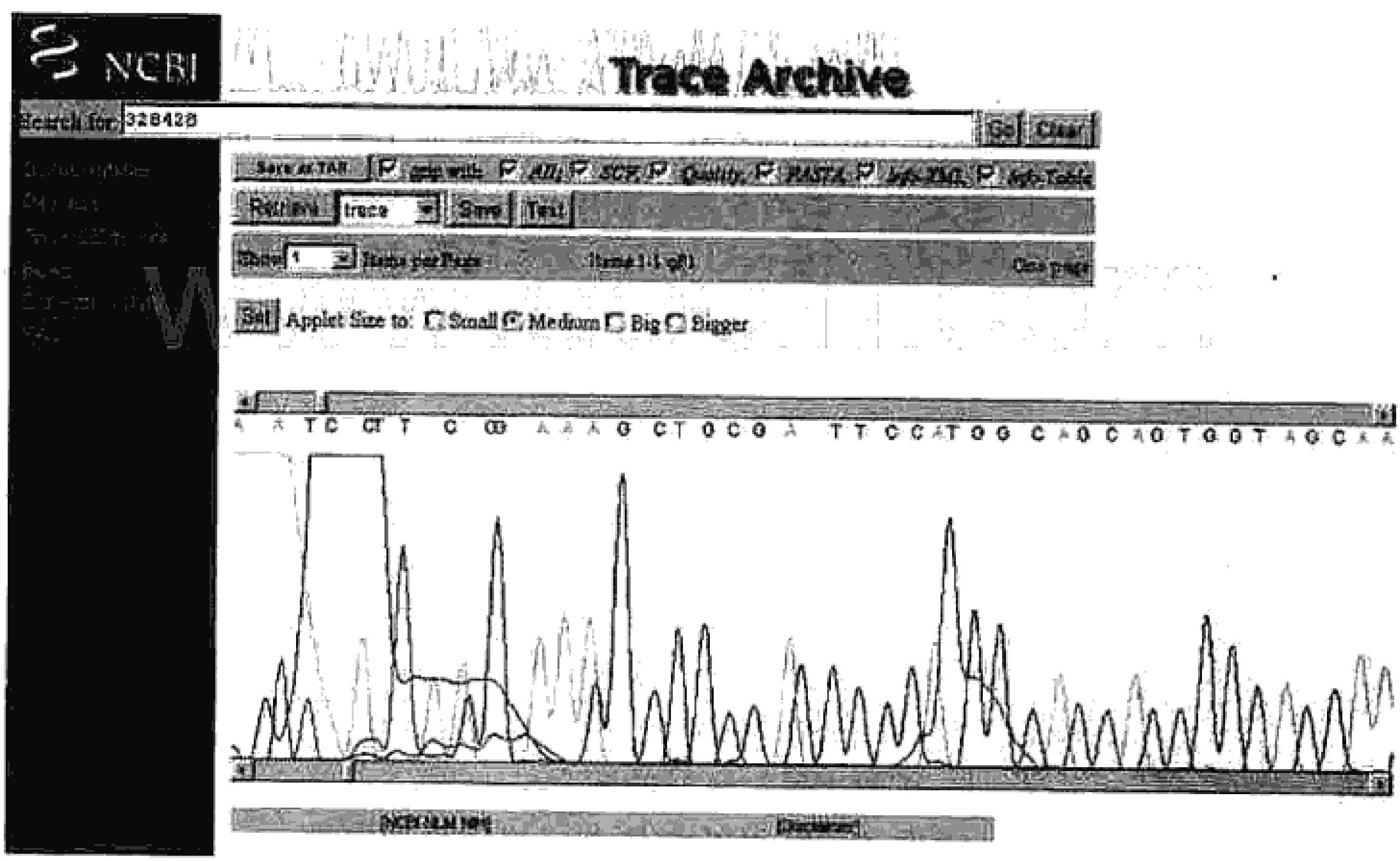


图 12-14 Trace 档案收集了多个基因组测序计划的原始数据。它可以从 NCBI 的网站（http://www.ncbi.nlm.nih.gov/Traces/trace.cgi）或者 EBI 的 Trace 服务器（http://trace.ensembl.org/）上访问获得。(a) 以小鼠 *RBP4* 进行 BLASTN 搜索，得到 11 个 Trace 档案匹配。出现图中的结果（一些在查询对象的 3′端，另一些在 5′端）是因为查询对象是一个表达的转录本，在 350bp 的对应位置上有一个内含子。(b) 点击第一个数据库匹配（ti 328428）可以得到 FASTA 格式的序列或者 DNA 测序反应的跟踪数据。根据这些跟踪数据便于估计基因组 DNA 记录的原始数据的质量。在一些情况下，染料终止反应（或者别的 DNA 测序技术）会产生不确定的结果，取得这些原始数据使得使用者可以对序列的质量做出有根据的判断

数据［图 12-14(b)］。

表 12-17 总结了 Trace 档案中的几种主要生物。

表 12-17 被选择生物的现存序列痕迹（截止到 2003 年 3 月）

生物体	俗名	序列痕迹的数目/个
冈比亚按蚊(*Anopheles gambiae*)	蚊子	4888159
黄牛(*Bos taurus*)	母牛	426683
线虫(*Caenorhabditis briggsae*)	蠕虫	2354917
斑马鱼(*Danio rerio*)	斑马鱼	12125895
家猫(*Felis catus*)	猫	460012
黄豆(*Glycine max*)	大豆	157399
人类(*Homo sapiens*)	人	18241156
小家鼠(*Mus musculus*)	小鼠	57591595
大家鼠(*Rattus norvegicus*)	大鼠	37781311
红鳍东方鲀(*Takifugu rubripes*)	河豚	1953192
黑青斑河豚(*Tetraodon nigroviridis*)	河豚	2688931
玉米(*Zea mays*)	谷物	69888

注：1. 这些数据信息包含了基于 WGS、EST 和 BAC 的项目。
2. 数据来自于 http://trace.ensembl.org/或者 http://www.ncbi.nlm.nih.gov/Traces/trace.cgi。

12.8.9 基因组注解：基因组 DNA 的特征

基因组测序完成后，我们得到了它的准确的大小，得到了完整（或者几乎完整）的核苷酸序列。基因组注解是调查基因组 DNA 的全貌、描述 DNA 的关键特征的一个过程（Stein，2001）。图 12-15 是 Ensembl 服务器给出的基因组注解的流水线的示例。

有关这个序列的特征，我们可能会提出以下三个基本问题。

① 总的 GC 含量或者其他核苷酸成分是多少？大多数真核基因组的 GC 含量界于 35%～45%之间，而细菌则具有一个宽得多的范围（图 12-16）。

② 有哪些重复的 DNA 序列，在什么地方？RepeatMasker 等程序可以识别并标记出重复的成分，比如 Alu 重复。我们将在第 15～第 17 章讨论这种重复。GLIMMER 和 GRAIL 等程序（见下文介绍）组合了多种可以识别基因组 DNA 中的重复成分的算法。

我们将在图 16-4～图 16-10 中演示几个重复 DNA 的例子以及识别和标记它的软件。

③ 现在一共有多少个已知基因（编码蛋白质的序列）？基因可以被一定数量的特征所确定，包括：

a. 基因的特殊密码子偏好；

b. 重复 DNA 序列的缺乏；

c. 启动子特异结合位点区域信号的出现。

这些特征在真核生物和原核生物的 DNA 序列之间是存在很大差异的。我们将会在第 14 章（重点关注细菌和太古细菌）和第 15～第 17 章（重点关注真核生物）中做更进一步的详细描述。

12.8.10 给细菌基因组做注释

细菌和太古细菌的基因组都包含附加的、相对较小的基因间区域。一般来说，这些基因组都是环状的，并且基因在 DNA 组序列中的分布密度大致为每 1000 个碱基对中含有一个

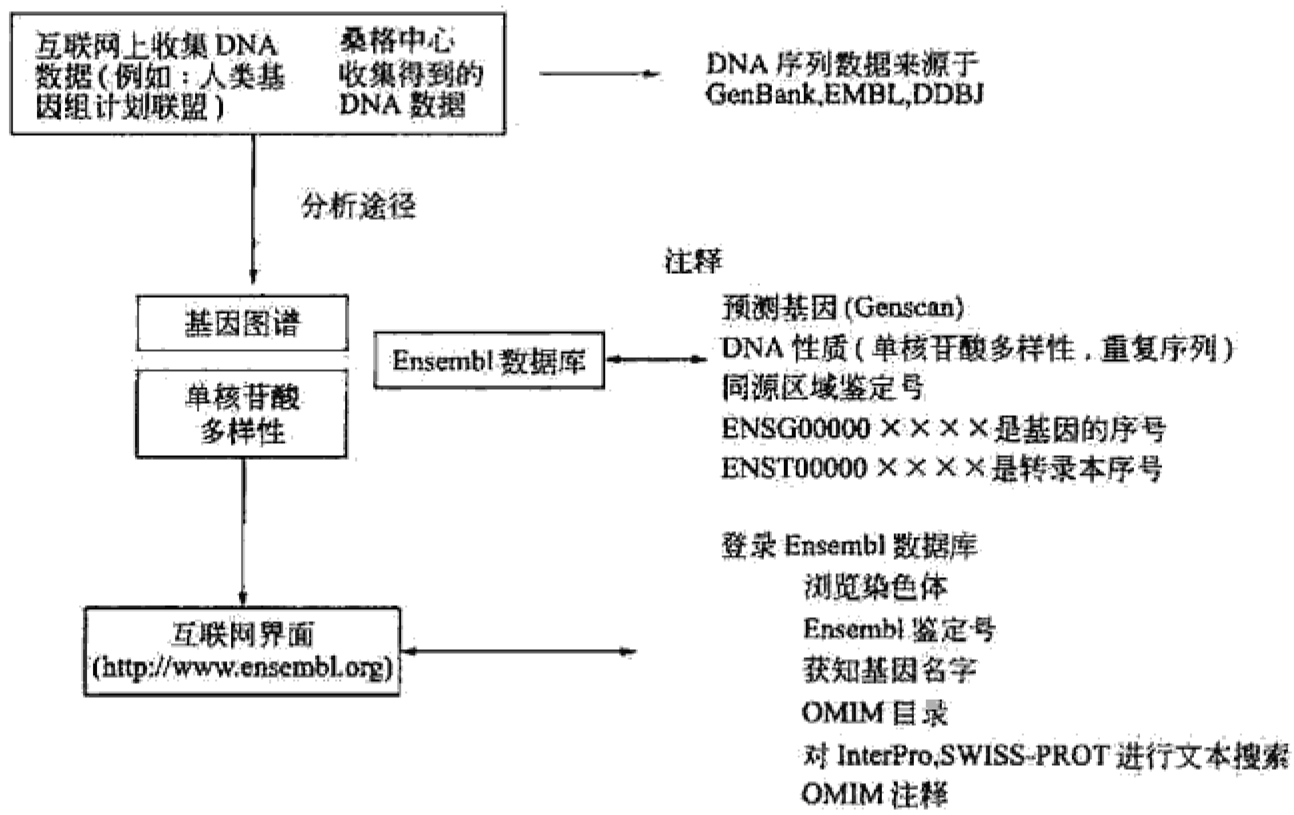


图 12-15 Ensembl 注解通道的简单综述（图示来源于 http://www.ensembl.org/Docs/enstour/）。Ensembl 是一个连接 EBI-EMBL 和 Sanger 研究所的工程计划，在这个计划中人们想要自动给那些已经被测序的 DNA 序列（例如人类基因组计划中得到的序列或者其他被测序的生物：小鼠，大鼠，斑马鱼，河豚，果蝇和两种线虫等）做注释或者标记

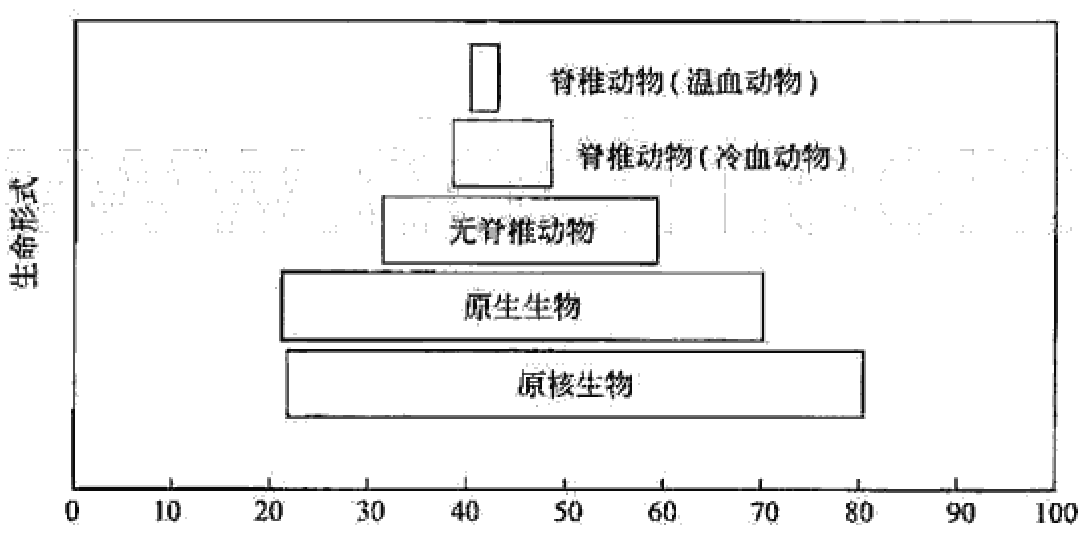


图 12-16 原核生物、原生生物、无脊椎动物和脊椎动物的鸟嘌呤十胞嘧啶（GC）含量分布。值得注意的是大部分真核生物基因组的 GC 含量为 40%～45%，而细菌和太古代的细菌却拥有一个更为广阔的 GC 含量范围。图示的结果来源于 Bernardi 和 Bernardi 在 1990 年发表的基于 20 世纪 70 年代和 80 年代研究成果的文献。目前的真核生物基因组测序计划（参见本书第 16 章中的详细描述）显示了不同生物体含有差异较大的 GC 含量值：19.4%（恶性疟原虫，*P. falciparum*）、22.2%（黏液霉菌，*Dictyostelium discoideum*）、34.9%（拟南芥，*A. thaliana*）、36%（线虫，*C. elegans*）、38.3%（啤酒酵母，*S. cerevisiae*）、41.1%（人类，*homo*）、42%（小鼠，*M. musculus*）和 43.3%（亚洲栽培稻，*O. sativa*）。对于那些已经完成测序的原核生物，GC 含量值的范围从 26%（溶尿尿原体，*Ureaplasma urealyticum parvum*）到 72%（链霉菌，*Streptomyces coelicolor*）（见表 14-3）。数据使用经授权

有效基因。对于原核生物来说，基因可以通过简单确定长开放式阅读框（ORF）的出现来确定，这些长开放式阅读框的长度一般都远远大于一些系统设定的阈值，如 90 个核苷酸的长度（30 个氨基酸的长度大概为 3ku 蛋白质序列的长度）。这样的程序有 GLIMMER 和 GenMark，能够有效确定基因在细菌基因组序列上的位置（详情参见 Baytaluk 等在 2002 年发表的相关的综述文章）（表 12-18）。

表 12-18 西南生物技术和生物信息中心总结的有关寻找新基因的相关程序（http://www.swbic;arg/links/1.4.3.2php）

资　源	来源/相关描述	超链接(URL)
CM Gene Finder(CM 新基因发现)	提供了一些工具来确定新基因	http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html
GeneFind(新基因发现)	蛋白质家族的鉴定问题和关键词搜索	http://pir.georgetown.edu/gfserver
GeneMark(基因标签)	提供了一些基因预测的程序	http://opal.biology.gatech.edu/GeneMark/
Gene Parser(基因分析)	在 DNA 基因序列中确定蛋白质编码区域	http://beagle.colorado.edu/～eesnyder/GeneParser.html
GeneScan(基因扫描)	脊椎动物 DNA 基因组中完整基因结构的确定	http://genes.mit.edu/GENSCAN.html
GLIMMER	在微生物 DNA 序列中寻找新基因的系统	http://www.tigr.org/software/glimmer/
GRAIL	基因识别和某些因特网链接的组合	http://compbio.ornl.gov/Grail-bin/EmptyGrailForm
ORF Finder(开放式阅读框寻找)	寻找开放式阅读框信息	http://www.ncbi.nlm.nih.gov/gorf/gorf.html
Procrustes	用可变剪切的方法来进行基因识别	http://www-hto.usc.edu/software/procrustes/

注：更多的相关程序清单可以在网站上得到。

Glimmer 是用来确定原核生物 DNA 序列上基因的一个程序。该程序需要两个输入元素：一个是 DNA 基因序列文件（FASTA 格式的文件），另一个是有关基因识别的马尔科夫模型。我们可以来看一个简单的 GLIMMER 程序输出的实例（详情参见图 14-7）。

12.8.11 真核基因的注释

与细菌基因组不同的是，真核基因组包含基因和大部分非编码的 DNA 序列。那些非编码区域包括一些具有调控功能的重复 DNA 的基因序列，以及那些从成熟的 RNA 转录物中分离出来的用于中断外显子的内含子部分。一个有关基因测序项目的主要焦点是确定基因组中的所有基因。但是，这就非常有必要给不同的基因形式下定义并且给出定义它们的相关准则，其中包括蛋白质编码基因、假基因和各种各样的 RNA 基因。本书将在第 15 章（关于真菌）和第 17 章（关于人类）中作详细介绍，并在第 6 章中对于真核生物新基因的确定作了特别详细的介绍。在本章中我们将会对如何在真核 DNA 基因组序列中确定基因的两种主要方法做一简单的讨论（图 12-17）。其中第一种方法是基于 DNA 基因组的表达序列标签（ESTs 或者 cDNAs 序列）的。因为 ESTs 序列是独立于 DNA 基因组序列获得的，所以这个方法也被称作“外部法”。全长的 cDNA 序列是通过实验证据从基因外显子的扩展序列定义中推导出来的，所以其可用性是无法准确估计的。作为选择，一个“内部”的方法是单单通过分析 DNA 基因组序列来预测基因的结构（外显子和内含子），寻找诸如开放式阅读框、外显子/内含子边界、启动和终止密码子，以及编码区域的密码子使用偏好等。关于基因预

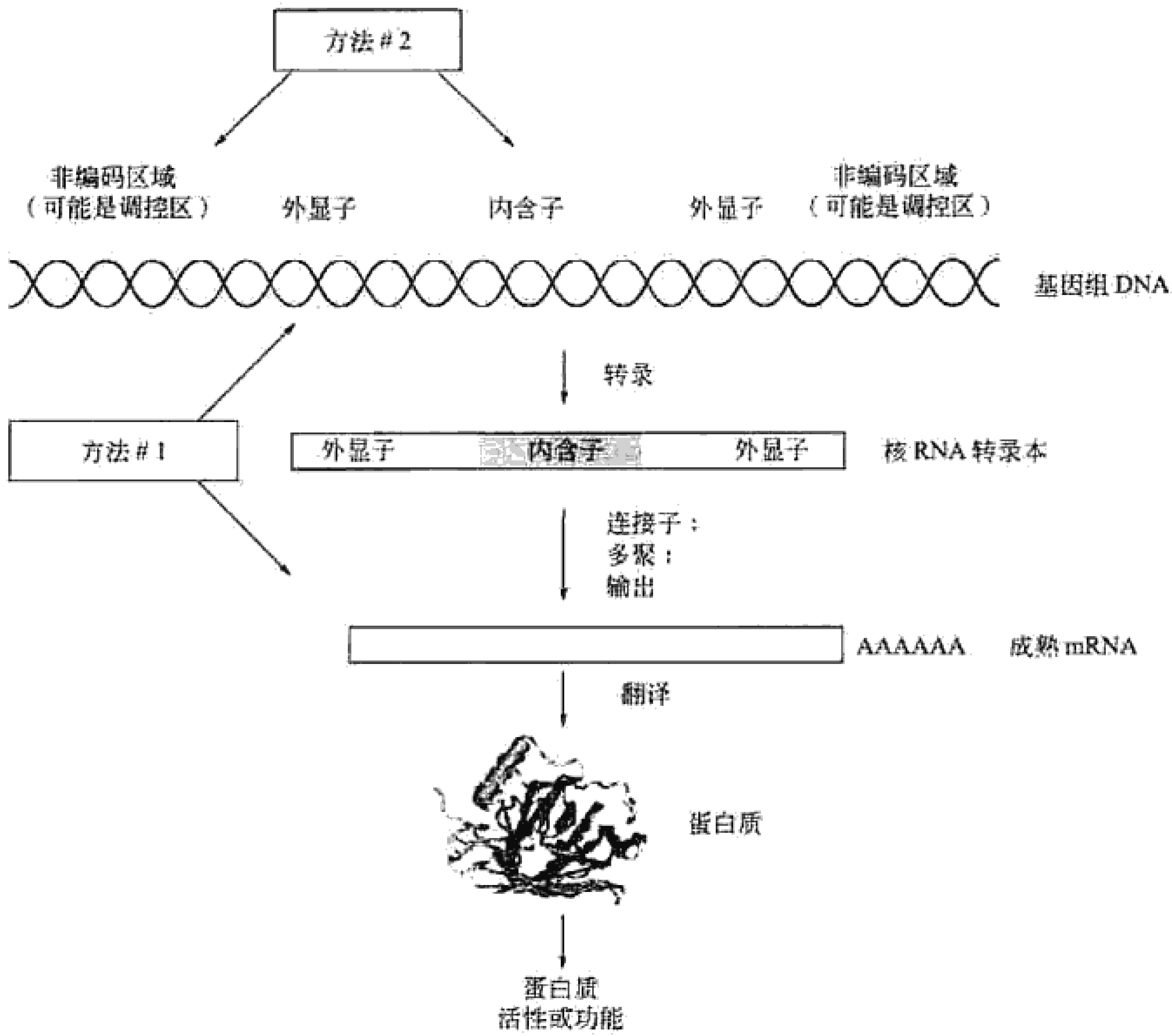


图 12-17 从DNA基因组序列中寻找蛋白质编码基因的两种主要方法。第一种方法也称作基于同源的（或称外部法）新基因寻找方法，是以表达序列标签与DNA基因组序列做比较来得到的。ESTs（表达序列标签）来源于同一个生物体RNA序列所产生的cDNA片段，不同生物体中可用的表达序列标签数目范围为从上千到上百万个不等。当一个表达序列标签与一段DNA基因序列相匹配时，就提供了一个确定蛋白质编码基因的有力证据。第二个方法又称为基于算法（或者内部法）的新基因寻找方法，着重于对DNA基因序列核苷酸组成特征的分析（如对长的开放式阅读框出现率的分析），例如在一个启动密码子和终止密码子之间存在一段长度满足至少300个氨基酸阈值的基因序列就是一个长的开放式阅读框出现的强信号之一。而内含子的出现（通常是在真核生物基因组中）将会使分析变得复杂化。编码区域的碱基组成通常与非编码区域存在着戏剧性的巨大差异，这个特征在第二种算法中也经常被作为设计寻找新基因算法的一个考虑因素

测的相关软件清单请参见图16-12～图16-18中的介绍。

12.8.12 总结：基因组测序计划中存在的问题

基因测序计划具体实行过程中遇到的一系列基本问题：

① 我们是否可以分别确定出蛋白质编码基因和RNA编码基因？

② 我们是否可以给它们指派相应的生物学功能？

③ 我们是否可以确定（或者预测）所有基因产物的结构？

④ 我们是否可以通过每个基因产物来重构转录调控网络和代谢调控网络？

⑤ 我们是否可以将基因型和表现型一一对应起来？例如，我们是否可以解释为什么人类对相同的致病菌或者环境毒素的敏感性差异如此巨大呢？我们是否可以解释为什么两种炭疽菌或者疱疹病毒的致病性差别如此大呢？

⑥ 我们是否可以定义生命的进化历史？其中部分内容可能可以通过对分子系统发生学和比较基因组学的研究来完成。这种方法被称作系统发生基因组学（图 12-18）（Eisen 和 Hanawalt，1999；Eisen 和 Fraser，2003）。

基因分析 ① 关注基因的相关数据库 ② 确定同源组 ③ 预测功能
基因组分析 ④ 不同种族中该基因的出现和缺失 ⑤ 修正出现和缺失值 ⑥ 预测表现型
进化分析 ⑦ 基因树 ⑧ 进化分布模式 ⑨ 一致性 ⑩ 基因进化事件 ⑪ 信号通路进化

图 12-18 系统发生基因组学描述了一个基于进化的基因组分析方法（数据来源于 Eisen 和 Hanawalt 1999 年的研究成果）。数据经授权使用

12.9 展望

从 1995 年开始我们进入了一个全新的生物学时代，在这个时代中，部分生物体的完整基因组序列被测定。很快，数以百计的生物体的完整序列的获得将变得可能。随着 2003 年人类基因组测序工作的完成，人类将进入“后基因组时代”。

基因组测序工作带动了系统发生分析学革命性的进展。目前的生命树版本包括 3 个主要分支（细菌、古细菌和真核生物）。在接下去的时间里，越来越多的分子序列数据将能够更清楚地解释关于地球上生命起源的一些关键问题：

① 行星上一共存在着多少个物种？

② 生命是如何从 4BYA 之前进化到现在的？

③ 为什么某些与致病菌亲缘关系相近的生物体是无毒性的？

④ 哪些是使人类和其他生物体致病的突变？

12.10 常见问题

当各个研究团体产生大量的 DNA 序列数据后，有关这些数据的解释将存在很多的缺陷。基因组测序本身存在着一个相关错误率（大概每 10000 个完成测序的 DNA 序列中存在一个核苷酸的错误）。这样，在计算 DNA 基因序列可能存在的多态性和突变时，评价序列数据的质量就变得很重要。即便序列信息是正确的，算法也不能很好地解决诸如如何找到真

核生物 DNA 序列中蛋白质编码基因之类的问题；在第 16 章中，我们将会对基因测序计划（例如大米和人类基因组的例子）做一详细介绍，在这些例子中研究者利用每一个基因组序列的修正值来戏剧性地改进外显子和基因的预测模型（对于细菌 DNA 来说，其普遍缺乏内含子，所以其准确率较高）。一旦蛋白质编码基因或者其他类型的基因被确定以后，有关基因组注释的工作存在着大量的错误率（Brenner，1999）。那么，关于仔细估计基因功能的注释就变得非常重要，并且有关基因功能的问题也必须用生物学和计算机的双重准则来衡量。

12.11 问题讨论

你能否决定优先追踪哪个基因组测序工程，比如猪的基因组测序工程？你是否能给生物体设定相互区分的优先顺序？

12.12 问题

图 12-1 展示了一棵基于 rRNA 序列的进化树。根据 3-磷酸甘油醛脱氢酶（GAPDH）蛋白序列来重建这棵进化树。这两棵进化树的相似性如何？用什么方法可以解决它们之间的区别？

12.13 自测题

1. 第一个被完整测序的基因组序列是

(a) 啤酒酵母（*Saccharomyces cerevisiae*）的 3 号染色体；

(b) 流感病毒（*Haemophilus influenzae*）；

(c) ϕX174；

(d) 人类的线粒体基因组

2. 一个普通真核生物的线粒体基因组一般编码多少个蛋白质（包括 RNAs）？

(a) 10；

(b) 100；

(c) 1000；

(d) 10000

3. 短语“全基因组鸟枪测序法”代表

(a) 是一个测定整个基因组序列的策略之一，指通过将 DNA 序列打碎和使用寡聚核苷酸作为引物来遍历整个 DNA 基因组序列的过程；

(b) 是一个测定整个基因组序列的策略之一，指通过将 DNA 序列打碎，构建一个克隆库，并且使用与一些已知的染色体位点（叠连群：contig）相对应的寡核苷酸引物来进行测序的过程；

(c) 是一个测定整个基因组序列的策略之一，指通过将 DNA 序列打碎，构建一个克隆库，确定库中每个大片段（叠连群：contig）的染色体位点，对其中的小片段进行测序，然后将这些片段重新装配为一张完整的基因图谱的过程；

(d) 是一个测定整个基因组序列的策略之一，指通过将 DNA 序列打碎，构建一个克隆库，对其中的小片段进行测序，然后将这些片段重新装配为一张完整的基因图谱的过程

4. 根据基因组序列来预测蛋白质编码基因算法的最大问题是

(a) 相关软件太难使用；

(b) 假阴性率太高，丢失了许多正确的外显子位点；

(c) 假阳性率太高，许多不是外显子的

序列部分被指定为外显子；

(d) 假阳性率太高，许多外显子的功能未知

5. 在主要的基因组测序中心里对 DNA 基因组进行程序的消耗大致为

(a) 每个碱基 1 美分；

(b) 每个碱基 1 美元；

(c) 每个碱基 10 美元；

(d) 每个基因组 1000 美元

6. 现在已经有数以百计的基因组序列被完成了，根据各种基因组的种属关系，它们中的大多数是

(a) 病毒的；

(b) 细菌的；

(c) 古细菌的；

(d) 细胞器的（线粒体和质粒）；

(e) 真核生物的

7. 应用于确定基因组序列的准则有很多，下列哪一个不是主要的准则之一？

(a) 项目消耗的多少；

(b) 是否与人类疾病相关；

(c) 基因组的大小；

(d) 基因组中所含基因的数目；

(e) 是否与最基本的生物学过程相关

参考文献

Anderson, S., et al. Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457-465 (1981).

Adams, M. D., et al. The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185-2195 (2000).

Alm, R. A., et al. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**, 176-180 (1999).

The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815 (2000).

Anderson, L. Identification of mitochondrial proteins and some of their precursors in two-dimensional electrophoretic maps of human cells. *Proc. Natl. Acad. Sci. USA* **78**, 2407-2411 (1981).

Andersson, S. G., et al. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133-140 (1998).

Baytaluk, M. V., Gelfand, M. S., and Mironov, A. A. Exact mapping of prokaryotic gene starts. *Brief Bioinform.* **3**, 181-194 (2002).

Benton, M. J., and Ayala, F. J. Dating the tree of life. *Science* **300**, 1698-1700 (2003).

Bernardi, G., and Bernardi, G. Compositional transitions in the nuclear genomes of cold-blooded vertebrates. *J. Mol. Evol.* **31**, 282-293 (1990).

Blackmore, S. Environment. Biodiversity update—progress in taxonomy. *Science* **298**, 365 (2002).

Blattner, F. R., et al. The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453-1474 (1997).

Brenner, S. E. Errors in genome annotation. *Trends Genet.* **15**, 132-133 (1999).

Bult, C. J., et al. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058-1073 (1996).

Carlton, J. M., et al. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* **419**, 512-519 (2002).

Casjens, S., et al. A bacterial genome in flux: The twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Mol. Microbiol.* **35**, 490-516 (2000).

Chambaud, I., et al. The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acids Res.* **29**, 2145-2153 (2001).

Cole, S. T., et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537-544 (1998).

Cole, S. T., et al. Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007-1011 (2001).

Dandekar, T., et al. Re-annotating the *Mycoplasma pneumoniae* genome sequence: Adding value, function and reading frames. *Nucleic Acids Res.* **28**, 3278-3288 (2000).

Deckert, G., et al. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**, 353-388 (1998).

Douglas, S., et al. The highly reduced genome of an enslaved algal nucleus. *Nature* **410**, 1091-1096 (2001).

Dunham, I., et al. The DNA sequence of human chromosome 22. *Nature* **402**, 489-495 (1999).

Eisen, J. A. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr. Opin. Genet. Dev.* **10**, 606-611 (2000).

Eisen, J. A., and Fraser, C. M. Phylogenomics: intersection of evolution and genomics. *Science* **300**, 1706-1707 (2003).

Eisen, J. A., and Hanawalt, P. C. A phylogenomic study of DNA repair genes, proteins, and processes. *Mutat. Res.* **435**, 171-213 (1999).

Fauron, C., Casper, M., Gao, Y., and Moore, B. The maize mitochondrial genome: Dynamic, yet functional. *Trends Genet.* **11**, 228-235 (1995).

Ferretti, J. J., et al. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc. Natl. Acad. Sci. USA* **98**, 4658-4663 (2001).

Fleischmann, R. D., et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512 (1995).

Fox, G. E., et al. The phylogeny of prokaryotes. *Science* **209**, 457-463 (1980).

Fraser, C. M., et al. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403 (1995).

Fraser, C. M., et al. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**, 580–586 (1997).

Fraser, C. M., et al. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**, 375–388 (1998).

Galibert, F., et al. The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* **293**, 668–672 (2001).

Gardner, M. J., et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).

The C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**, 2012–2018 (1998).

Giaever, G., et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).

Glass, J. I., et al. The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. *Nature* **407**, 757–762 (2000).

Goffeau, A., et al. Life with 6000 genes. *Science* **274**, 546, 563–577 (1996).

Gordon, D., Desmarais, C., and Green, P. Automated finishing with autofinish. *Genome Res.* **11**, 614–625 (2001).

Graur, D., and Li, W.-H. *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA, 2000.

Haeckel, E. *The Evolution of Man: A Popular Expositon of the Principal Points of Human Ontogeny and Phylogeny*. D. Appleton and Company, New York, 1879.

Hattori, M., et al. The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* **405**, 311–319 (2000).

Hayashi, T., et al. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* **8**, 11–22 (2001).

Hedges, S. B., et al. Genomic timescale for the origin of eukaryotes. *BMC Evol. Biol.* **1**, 1–10 (2001).

Heidelberg, J. F., et al. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406**, 477–483 (2000).

Heilig, R. *Pyrococcus abyssi* genome sequence: insights into archaeal chromosome structure and evolution. Unpublished GenBank entry NC_000868 (http://www.ncbi.nlm.nih.gov), 2001.

Himmelreich, R., et al. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24**, 4420–4449 (1996).

Holt, R. A., et al. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129–149 (2002).

International Human Genome Sequence Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

Joyce, G. F. The antiquity of RNA-based evolution. *Nature* **418**, 214–221 (2002).

Kaiser, D. Building a multicellular organism. *Annu. Rev. Genet.* **35**, 103–123 (2001).

Kalman, S., et al. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nat. Genet.* **21**, 385–389 (1999).

Kaneko, T., et al. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3**, 109–136 (1996).

Katinka, M. D., et al. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414**, 450–453 (2001).

Kawarabayasi, Y., et al. Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain 7. *DNA Res.* **8**, 123–140 (2001).

Kawarabayasi, Y., et al. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res.* **5**, 55–76 (1998).

Kawashima, T., et al. Determination of the complete genomic DNA sequence of *Thermoplasma volvanium* GSS1. *Proc. Jpn. Acad.* **75**, 213–218 (1999).

Klenk, H. P., et al. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**, 364–370 (1997).

Knapp, S., et al. Taxonomy needs evolution, not revolution. *Nature* **419**, 559 (2002).

Koonin, E. V. Genome sequences: Genome sequence of a model prokaryote. *Curr. Biol.* **7**, R656–659 (1997).

Koonin, E. V. Computational genomics. *Curr. Biol.* **11**, R155–158 (2001).

Korab-Laskowska, M., et al. The Organelle Genome Database Project (GOBASE). *Nucleic Acids Res.* **26**, 138–144 (1998).

Kumar, S. and Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).

Kunst, F., et al. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256 (1997).

Lander, E. S., and Waterman, M. S. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2**, 231–239 (1988).

Lang, B. F., Gray, M. W., and Burger, G. Mitochondrial genome evolution and the origin of eukaryotes. *Annu. Rev. Genet.* **33**, 351–397 (1999).

Lilly, J. W., and Havey, M. J. Small, repetitive DNAs contribute significantly to the expanded mitochondrial genome of cucumber. *Genetics* **159**, 317–328 (2001).

Mayr, E. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. Belknap Harvard, Cambridge, MA, 1982.

May, B. J., et al. Complete genomic sequence of *Pasteurella multocida*, Pm70. *Proc. Natl. Acad. Sci. USA* **98**, 3460–3465 (2001).

Nelson, K. E., et al. Evidence for lateral gene transfer between *Archaea* and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329 (1999).

Ng, W. V., et al. Genome sequence of *Halobacterium* species NRC-1. *Proc. Natl. Acad. Sci. USA* **97**, 12176–12181 (2000).

Nierman, W. C., et al. Complete genome sequence of *Caulobacter crescentus*. *Proc. Natl. Acad. Sci. USA* **98**, 4136–4141 (2001).

Ohyama, K., et al. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature* **322**, 572–574 (1986).

Oliver, S. G., et al. The complete DNA sequence of yeast chromosome III. *Nature* **357**, 38–46 (1992).

Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).

Parkhill, J., et al. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**, 665–668 (2000a).

Parkhill, J., et al. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404**, 502–506 (2000b).

Pennisi, E. Taxonomy. Linnaeus's last stand? *Science* **291,** 2304–2307 (2001).

Perna, N. T., et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409,** 529–533 (2001).

Read, T. D., et al. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.* **28,** 1397–1406 (2000).

Ruepp, A., et al. The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature* **407,** 508–513 (2000).

Sanger, F., et al. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265,** 687–895 (1977).

She, Q., et al. The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. *Proc. Natl. Acad. Sci. USA* **98,** 7835–7840 (2001).

Shinozaki, K. M., et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J.* **5,** 2043–2049 (1986).

Simpson, A. J., et al. The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature* **406,** 151–157 (2000).

Skaletsky, H., et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423,** 825–837 (2003).

Smith, D. R., et al. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: Functional analysis and comparative genomics. *J. Bacteriol.* **179,** 7135–7155 (1997).

Stein, L. Genome annotation: From sequence to biology. *Nat. Rev. Genet.* **2,** 493–503 (2001).

Stephens, R. S., et al. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282,** 754–759 (1998).

Stover, C. K., et al. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406,** 959–964 (2000).

Takami, H., et al. Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. *Nucleic Acids Res.* **28,** 4317–4331 (2000).

Tettelin, H., Radune, D., Kasif, S., Khouri, H., and Salzberg, S. L. Optimized multiplex PCR: Efficiently closing a whole-genome shotgun sequencing project. *Genomics* **62,** 500–507 (1999).

Tettelin, H., et al. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287,** 1809–1815 (2000).

Tettelin, H., et al. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* **293,** 498–506 (2001).

Tomb, J. F., et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388,** 539–547 (1997).

Van de Peer, Y., De Rijk, P., Wuyts, J., Winkelmans, T., and De Wachter, R. The European small subunit ribosomal RNA database. *Nucleic Acids Res.* **28,** 175–176 (2000).

Venter, J. C., et al. The sequence of the human genome. *Science* **291,** 1304–1351 (2001).

Weber, J. L., and Myers, E. W. Human whole-genome shotgun sequencing. *Genome Res.* **7,** 401–409 (1997).

White, O., et al. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286,** 1571–1577 (1999).

Wilson, E. O. *The Diversity Of Life.* W. W. Norton, New York, 1992.

Woese, C. R. Default taxonomy: Ernst Mayr's view of the microbial world. *Proc. Natl. Acad. Sci. USA* **95,** 11043–11046 (1998).

Woese, C. R., Kandler, O., and Wheelis, M. L. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **87,** 4576–4579 (1990).

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., and Koonin, E. V. Genome trees and the tree of life. *Trends Genet.* **18,** 472–479 (2002).

Wood, V., et al. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415,** 871–880 (2002).

Wuyts, J., Van de Peer, Y., Winkelmans, T., and De Wachter, R. The European database on small subunit ribosomal RNA. *Nucleic Acids Res.* **30,** 183–185 (2002).



严重急性呼吸道综合征（SARS）于 2002 年底在中国出现并迅速蔓延成为一个全球性的传染病。它是由一类冠状病毒（如图所示）引起的。SARS 病毒的基因组（编号 NC _ 0047718）包含 29727 个核苷酸，它的测序工作是由加拿大和美国疾病控制预防中心（CDC）的几个小组于几周内分别独立完成的。这里所列的蛋白质序列（编号 NP _ 828849）含有 7073 个氨基酸残基（预测其分子量约为 790000）。通过在 Entrez 蛋白质数据库查询“790000：999999 [MolWt]”可发现目前已知的分子量大于或等于 790000 的蛋白质仅有 100 种左右。图中的冠状病毒图像来自 CDC 的网页（http://phil.cdc.gov/Phil/detail.asp?id=3492）

第 13 章

已完成测序的基因组：病毒基因组

13.1 引言

在本章中我们将会考察如何运用生物信息学方法对病毒进行研究。病毒是小的、具有感染性的、必须在细胞内生存的寄生生物。它们依赖宿主细胞进行复制。病毒体（病毒颗粒）由一个核酸基因组以及封装这个基因组的蛋白质外壳（衣壳体）构成，其外侧还可能有来源于宿主细胞的并布满病毒糖蛋白的脂双层包被。与其他生物的基因组不同，病毒的基因组只能是 DNA，或者只能是 RNA。此外，它们可能是单链、双链或部分双链，而且可能是环形、线形或分为多段（不同的基因分布在不同的核酸片段上）。

病毒缺乏独立生存所必需的生物化学机制。这是病毒和能够自由独立生存的生物体之间的基本区别。但是，由于它们尚具备自我复制和进化的特征，一直以来病毒在“生物”这个定义所涵盖的范围内处于一个边缘的位置。最大的病毒（例如痘病毒）的基因组碱基对数目可达几十万，几乎与最小的古细菌或细菌相当（例如 *Nanoarchaeum equitans* 和 *Mycoplasmagenitalium*）（第 14 章）。因此，下列现象并不只是巧合而已：最小的原核生物的基因组可能来源于更小的、有感染性的、必须在细胞内生存的个体（见第 14 章）。

地球上可能生存着几亿或几十亿种细菌和古细菌，然而目前已知的病毒却只有数千种。这种数量上的巨大差异或许反映了它们入侵宿主的不同特异需求。病毒感染所有形式的生命，包括细菌、古细菌以及真菌、植物、人类等各种真核生物。

虽然病毒是一种相对简单的个体，但是与另外两种致病体——类病毒和朊病毒相比还是稍显复杂。类病毒是一类通常引起植物疾病的、由大约 200～400 个核苷酸组成的环形 RNA 分子（Flores，2001）。类病毒微小的基因组并不编码任何蛋白质，而且其 RNA 分子本身具有酶的催化活性。朊病毒是具有感染性的蛋白质分子（Prusiner，1998）。Cruetzfeld-Jakob 疾病是最常见的人类朊病毒疾病（Johnson 和 Gibbs，1998）。在全世界范围内这种疾病发生的概率大约是 $1/10^6$，并且通常呈现为痴呆的症状。疯羊病和疯牛病是动物中的最常见的朊病毒疾病。

13.2 病毒的分类

在序列测定之前，形态学是病毒分类一个重要的依据。自 1959 年以来，电子显微术已经被用于描述超过 5100 个噬菌体（Ackermann，2001）以及其他入侵植物和动物的病毒的结构。96%的噬菌体是具有尾部结构的病毒，其余的呈杆状，二十面体状，或可变形状。许多病毒的电子显微镜图像可以在 ICTVdb（the database of the International Committee on Taxonomy of viruses，国际病毒分类学委员会）数据库中获得（Buchen-Osmond，1997）。图 13-1 中展示的就是几张从 ICTVdb 中获得的病毒的电子显微镜图像。

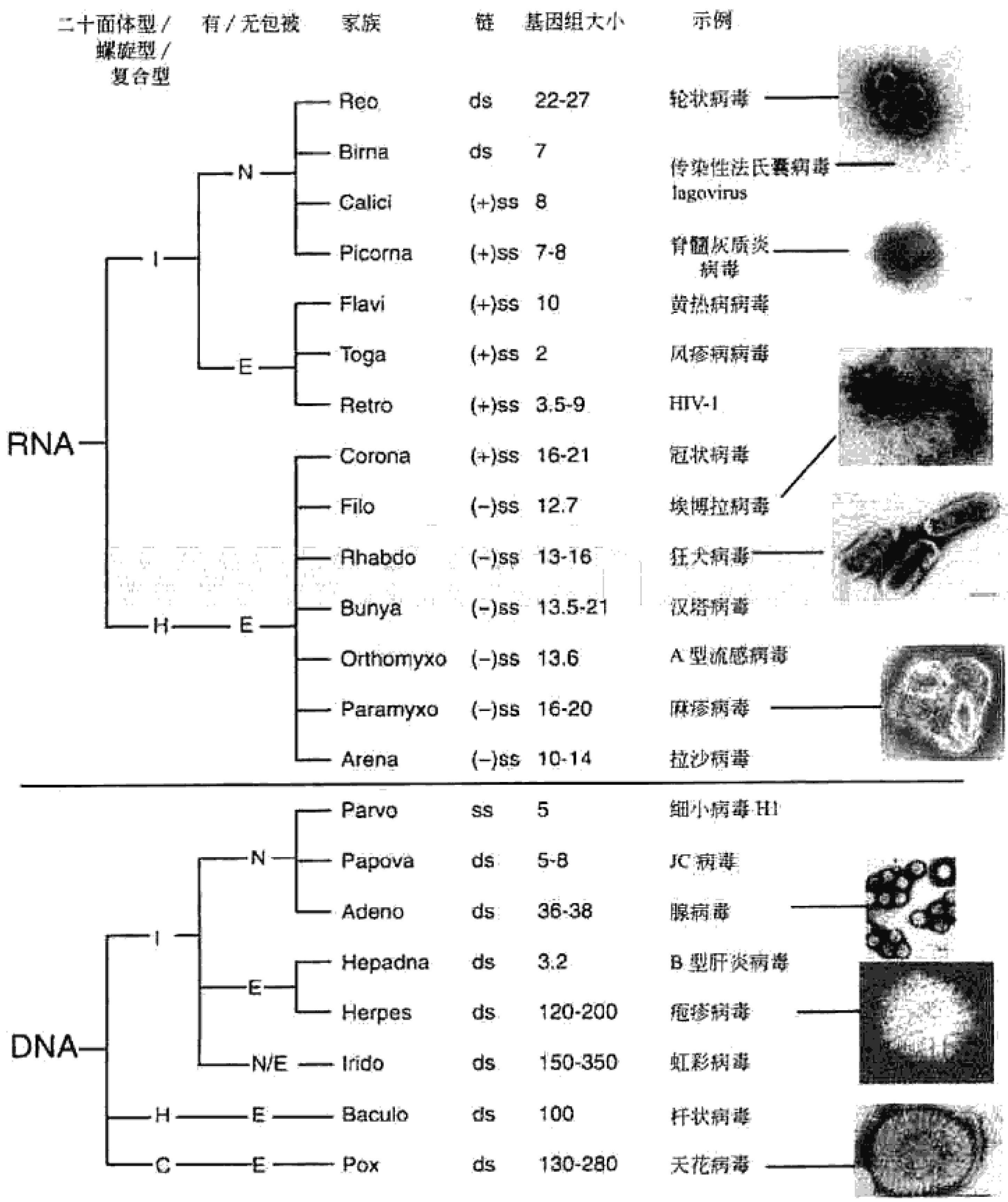


图 13-1 病毒的分类［根据 ICTVdb 和 Flint 等（2000）改写。电子显微镜照片来自 ICTV 网站］

ICTV 网站位于 http://www.ncbi.nlm.nih.gov/ICTVdb/index.htm，由 Cornelia Buchen-Osmond（澳大利亚国家大学生物信息学研究组）建立微病毒的 GenBank 编号是 AABV 01000000。

另一个对病毒进行分类的基本依据是病毒颗粒内包被的核酸基因组的类型。病毒体包含 DNA 或 RNA；核酸可能是单链的或双链的，而且翻译可能从正义链、反义链或两者同时发生，这些都可以是分类的标准。双链基因组以 DNA 或 RNA 双链中的一条链作为模板合成子链。单链 DNA 病毒或 RNA 病毒以它们自身作为模板合成互补链。复制过程中可能包括形成稳定或短暂双链中间物的过程。一些单链 RNA 基因组会首先利用逆转录酶（RNA-依赖性脱氧核糖核酸合成酶）将自身 RNA 转录为 DNA。例如在 HIV-1 中，*pol* 基因编码逆转录酶。

ICTV 定期召开会议以提出一个得到广泛接受的病毒分类标准。ICTV 认识目、科、属和种的分类单元。ICTV 数据库（第六次报告）将病毒进一步细分为约 71 个科，9 个亚科和 164 个种类［Mayo 和 Pringle（1998）概述］。图 13-2 所示是 ICTV 网站上的病毒在线描述的一个例子。

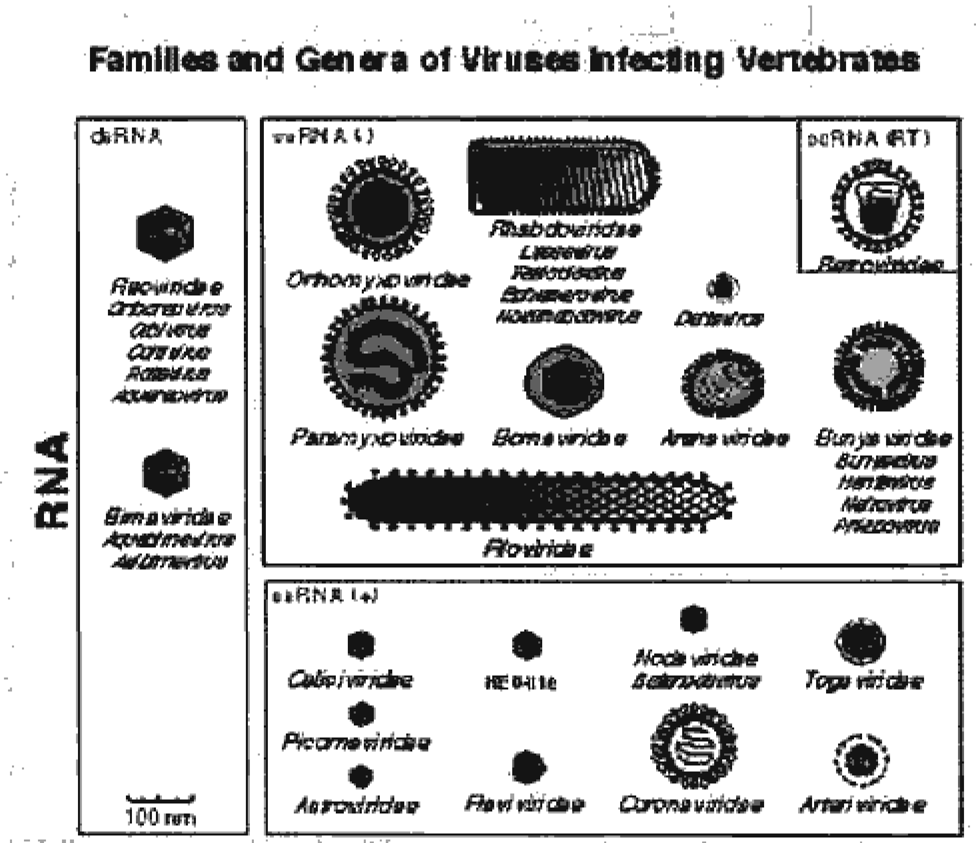


图 13-2 ICTV 数据库中示意病毒多态性的图例（参见 http://www.ncbi.nlm.nih.gov/ICTV/Diagrams/Verrna.html）

美国国家卫生研究院（NIH）下属的过敏和传染病国家研究所（NIAID）提供病毒及其他疾病的有关信息，网址为 http://www.niaid.nih.gov/publications/。

图 13-1 和表 13-1 列出了一些主要的病毒种类。它们的基因组大小差别很大，非常小的病毒（例如麻疹病毒）的基因组只有 2000 个碱基对而一些大的病毒的基因组的大小超过 35 万个碱基对。已经发现了一种巨大的病毒（mimivirus），它的双链环形基因组含有 80 万个碱基对（La Scola 等，2003）。它的成熟粒子直径达到 400nm，其尺寸甚至超过了许多细菌和古细菌。

另一种与上述完全不同的方法以是否引起人类疾病以及引起哪些疾病为依据对病毒进行分类。许多病毒性疾病能通过接种疫苗（表 13-2）来预防。其他病毒（如天花病毒）由于近期存在被恐怖分子利用的潜在可能性而受到关注（Cieslak 等，2002）。由天花病毒所引起的天花疾病于 1977 年被宣布在世界范围内被消灭了，而在美国，常规接种天花疫苗早在 1972 年就已经停止。

表 13-1 根据核酸组成对病毒进行分类

核酸	链	家族	示例	编号	碱基对数量
RNA	单链	小核糖核酸病毒科	I 型人脊髓灰质炎病毒	NC_002058	7440
		囊膜病毒科	风疹病毒	NC_001545	9755
		黄热病病毒科	黄热病病毒	NC_002031	10862
		冠状病毒科	冠状病毒	NC_002645	27317
		棒状病毒科	狂犬病毒	NC_001542	11932
		副黏病毒科	麻疹病毒	NC_001498	15894
		正黏病毒科	A 型流感病毒(节段 1)	NC_002023	13585
		布尼亚病毒科	汉塔病毒	—	—
		沙粒病毒科	拉沙热病毒	J04324	3402
		逆转录病毒科	HIV	NC_001802	9181
	双链	呼肠孤病毒科	轮状病毒		
DNA	Single	细小病毒科	H1 型细小病毒	NC_001358	5176
	Mixed	嗜肝病毒科	B 型肝炎病毒	NC_001707	3215
	Double	乳多空病毒科	JC 病毒	NC_001699	5130
		腺病毒科	人类腺病毒亚型 17	NC_002067	35100
		疱疹病毒科	I 型人类疱疹病毒	NC_001806	152261
		痘病毒科	天花病毒	NC_001559	191737

注：部分改编自 Schaechter 等（1999，第 292 页），已获授权。

表 13-2 可通过疫苗预防的病毒

疾病	病毒	描述
A 型肝炎	A 型肝炎病毒	引发肝炎
B 型肝炎	B 型肝炎病毒	引发肝炎
流感	A 型或 B 型流感病毒	在美国每年导致 20000 人死亡
麻疹	麻疹病毒	见后文
腮腺炎	rubulavirus	一种淋巴疾病
小儿麻痹症	脊髓灰质炎病毒	一种发生在脊髓灰质中的炎症，导致炎症部位神经元死亡
轮状病毒病	轮状病毒	幼儿痢疾的最常见致病源；全世界每年有 600000 名儿童死于这种疾病
风疹	风疹病毒	也叫德国风疹
天花	天花病毒	1977 年被消灭
水痘	带状水痘病毒	75%的小于 15 岁儿童都患有这种疾病

注：改编自 http://www.cdc.gov/nip/diseases/disease-chart-hcp.htm。

13.3 运用生物信息学方法研究病毒学问题

生物信息学工具非常适合于解决一些尚未获得答案的病毒学问题。

① 为什么艾滋病病毒选择性地侵染某一物种（人类），而与之具有紧密关联的另一病毒（猿的免疫缺陷病毒）只侵染猿猴而不侵染人类？对病毒及其宿主细胞受体进行序列分析有可能解决这个问题。

② 一些病毒为什么能够改换它们的天然宿主？1997 年，一种禽流感病毒感染了 18 个人并导致其中 6 人死亡。对从这种感染人类的流感病毒中得到的基因进行序列分析，发现这些基因似乎并不决定其病毒的宿主特异性（Suarez 等，1998）。

③ 为什么某些病毒株比其他病毒株具有更高的致命性？1918 年的一次世界范围的流感暴发造成全球约 4000 万至 1 亿人死亡。研究人员通过对收集的 1918 年病毒样品和现代的流感病毒进行序列比较来研究这一问题（Reid 等，2002）。

④ 病毒躲避宿主免疫系统的机制是什么？我们将会在下面介绍某些疱疹病毒是如何获得人类免疫系统分子的同源性从而干扰人类反病毒的机制的。

⑤ 病毒从何处起源？对于这个问题有三个主要的理论。

a. 倒退理论认为病毒起源于那些更加复杂的细胞内寄生生物。这些寄生生物在去除了许多非本质特征后成为病毒。

b. 病毒可能起源于正常细胞成分的自我复制。

c. 病毒可能与它们的宿主细胞共同进化，起源于自我复制的 RNA 分子。

系统发育分析可以帮助确证这些理论。

⑥ 哪些疫苗最有可能是有效的？为具有大量序列多样性的病毒开发疫苗有两种主要的方法。第一种方法主要针对某一地方性传播的病毒子类。第二种方法首先从现有病毒序列推测出一个祖先序列或简并序列（consensus sequence），再以这个序列作为抗原进行疫苗开发（Gaschen 等，2002）。这些方法均依赖于分子的系统演化。

13.3.1 多样性和病毒的进化

对生物的分类应该体现其系统演化。对病毒来说，由于它们独特的难于理解的而且有时脆弱的本质，对它们进化历程的追踪很难达到像对古细菌、细菌和真核生物那样综合全面的程度。与活的生命一样，病毒受自身突变（遗传物质的变异）和自然选择的影响。但是病毒基因组的进化速度远远超过其他细胞的基因组，因而给当前对病毒进化的研究造成了困难。

病毒很难在考古学或历史学的样品中保存下来。有相当多的证据证明，早在一万年以前病毒就已经存在，这些证据包括人类的骨骼残骸、历史记录和其他历史遗物。然而，远古病毒的 DNA 或 RNA 还没有被找到。

RNA 病毒其基因组的 RNA 聚合酶一般缺乏校正能力。这导致其基因组的突变率比 DNA 基因组高 100 万～1000 万倍（McClure，2000）。对于 DNA 病毒，其突变率一般比宿主细胞高 10～100 倍。

除了高突变率外，许多病毒的复制速度也是极其惊人的。单个细胞能生产 10000 个脊髓灰质炎病毒颗粒，而一个被艾滋病病毒感染的个体一天能生产 10 亿个病毒颗粒。这种高突变率和高复制速度能导致病毒“类种”的形成（一群相关但并不完全相同的病毒）。

根据 George Gaylord Simpson（1963）的定义，“物种是实际的或潜在的，与其他物种存在生殖隔离的种群集合。一个进化中的物种是各物种共同祖先的后代，但有着自己独特的进化角色和倾向，从而从其他共同祖先的后代中分离出来，与其他后代有着不同的进化历程。”

许多病毒的基因组由相对独立的多个片段组成。这些片段能够在病毒复制过程中随机重组，从而在子代病毒中产生大量互不相同的子类（见下文关于艾滋病病毒的部分）。感冒病毒几乎每年都能引起大范围疾病流行就是这种原理的一个表现。

病毒时常处于强大的选择压力下，如宿主的免疫反应或抗病毒药物治疗。因此，艾滋病病毒快速的突变和复制确保某些病毒株可能通过突变产生对抗病毒药物的抗性，而且这些病毒将会经受环境的选择而存活下来。

病毒通过漫长的进化历程已经能够侵入系统发生树中所有物种：古细菌、细菌和真核生物。植物病毒（如番茄丛矮病毒）、动物病毒（如 SV40 病毒、鼻病毒和脊髓灰质炎病毒），以及噬菌体（如噬菌体 ΦX174）的衣壳蛋白中都具有“β-折叠桶”或“果冻卷”折叠结构（Hendrix，1999）。除非发生了显著的趋同进化，否则这种现象一般说明这些病毒是同源的。感染植物和动物的逆转录病毒具有双链 RNA 基因组以及封装它的特殊衣壳体。有一类噬菌体（例如 Φ6）也具有这些特征，这再一次说明感染不同物种的病毒之间具有同源性（Hendrix，1999）。非常值得注意的是，在对这些病毒基因组的基因以及（或

者）蛋白质的分析中并没有发现序列相似性，这再次凸显了病毒基因组高速进化的特点。

病毒基因组的高度多样性使我们无法根据其序列数据绘制出涵盖所有病毒的全面完整的系统发生树。这反映了病毒基因组形成历程中复杂的分子进化事件（McClure，2000）。

若干个针对某一病毒家族的系统发生树已经被绘制出来。在证实病毒的进化、宿主特异性、毒性以及其他生物学性质方面，这些系统发生树所提供的信息是不可或缺的。下面，我们将深入考察疱疹病毒的系统发生重建。艾滋病病毒（见下文）以及麻疹病毒、肝炎病毒等其他病毒的系统发生树也已经绘出。

13.3.2 从系统发生到基因表达：运用生物信息学方法研究疱疹病毒

疱疹病毒是一群包括单纯性疱疹病毒、巨细胞病毒和爱泼斯坦-巴尔病毒的双链 DNA 病毒家族。作为病毒种系发生的一个例子，McGeoch 等（1995）分析了疱疹病毒的 8 个保守基因并据此推测其种系发生。他们重建的种系发生与三类主要疱疹病毒——α-疱疹病毒、β-疱疹病毒和 γ-疱疹病毒的分类法相一致（图 13-3）。相似的分析（Davison，2002）提供了对疱疹病毒起源、功能和多样性的有益理解。每种疱疹病毒都与一种宿主生物相关联（虽然包括人类在内的一些宿主能够被多种疱疹病毒感染）。这种特异性说明疱疹病毒已经在数百万年来和它们的宿主共同进化。McGeoch 等（1995）估计，假定存在一个恒定的分子时钟，则疱疹病毒在公元前 2 亿 2000 万年至公元前 1 亿 8000 万年间出现，它的三个主要子谱系在

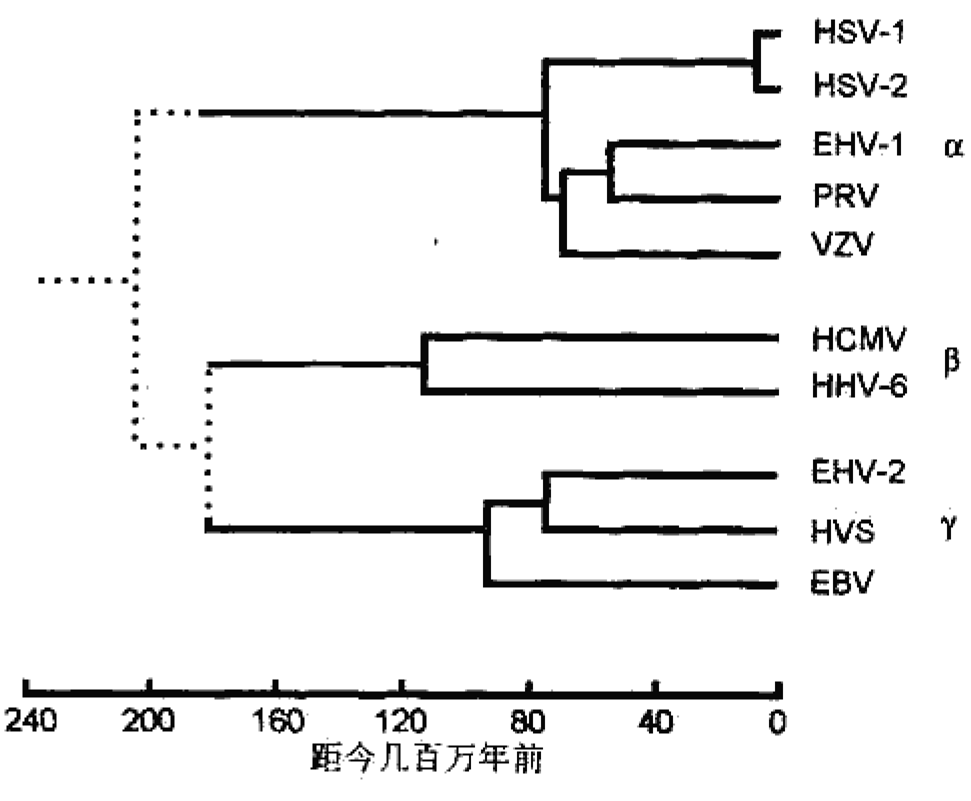


图 13-3 疱疹病毒的系统演化。α-疱疹病毒包括单纯性疱疹病毒亚型 1 和亚型 2（HSV-1 和 HSV-2，都以人类为宿主）、马疱疹病毒亚型 1（EHV-1）、伪狂犬病病毒（PRV，以猪为宿主）和带状水痘疱疹病毒（VZV，以人类为宿主）。β-疱疹病毒包括人类细胞巨化病毒（HCMV）和人类疱疹病毒亚型 6（HHV-6）。γ-疱疹病毒包括马疱疹病毒亚型 2（EHV-2）、松鼠猴疱疹病毒（HVS，以南美洲松鼠猴为宿主）、爱泼斯坦-巴尔病毒（EBV，以人类为宿主）和人类疱疹病毒亚型 8（HHV-8，以人类为宿主，图中没有表示）。系统发生树是运用 UPGMA 方法（第 11 章）对各病毒中一种保守的病毒糖蛋白进行多序列比对生成的。虚线表示可能的早期分化途径。图片来自 McGeoch 等（1995），经授权使用

公元前 8000 万年至公元前 6000 万年产生，大约与哺乳动物暴发（恐龙时代结束）同时。

卡波西瘤是最普遍的与艾滋病相关的肿瘤，是一种通常最早出现在皮肤上的恶性血管瘤。

接下来考察一种 γ-疱疹病毒——人类疱疹病毒亚型 8（HHV-8）。HHV-8 也被称为卡波西瘤关联疱疹病毒，它最初通过表现度差值分析在艾滋病患者的卡波西瘤损伤组织中被识别（Zhang 等，1994）。HHV-8 引起艾滋病关联的卡波西瘤和其他的病变，例如原发性流出物淋巴瘤和多中心卡瑟尔曼症。HHV-8 与恒河猴棒状病毒（RRV）非常类似，两者的区别可能与人类和恒河猴的物种差异相一致（Davison，2002）。黑猩猩体内既存在 HHV-8 也存在另一种与 HHV-8 相关联的病毒，这意味着这另外的一种病毒也可能感染人类。

HHV-8 具有潜伏期和烈性感染期这样周期循环的分子基础是什么？HHV-8 病毒的基因组大约有 140000 对碱基（NC _ 003409），而且编码超过 80 种蛋白质（Russo 等，1996）。我们能通过 NCBI 网站使用 Entrez 基因组工具（第 12 章）来探究这个基因组。NCBI 上还有用于病毒研究的其他资源。从一个病毒基因组的首页（图 13-4）你能链接到其他双链 DNA 病毒（如疱疹病毒），或者你能选用相关病毒的蛋白质聚类工具（CRP）（图 13-5）。从这个站点能浏览双链 DNA 病毒（图 13-6），并且能获得数十个疱疹病毒的基因组（图 13-7）。选择 HHV-8，能看到一个图片形式（图 13-8）或列表形式的（图 13-9）开放阅读框。

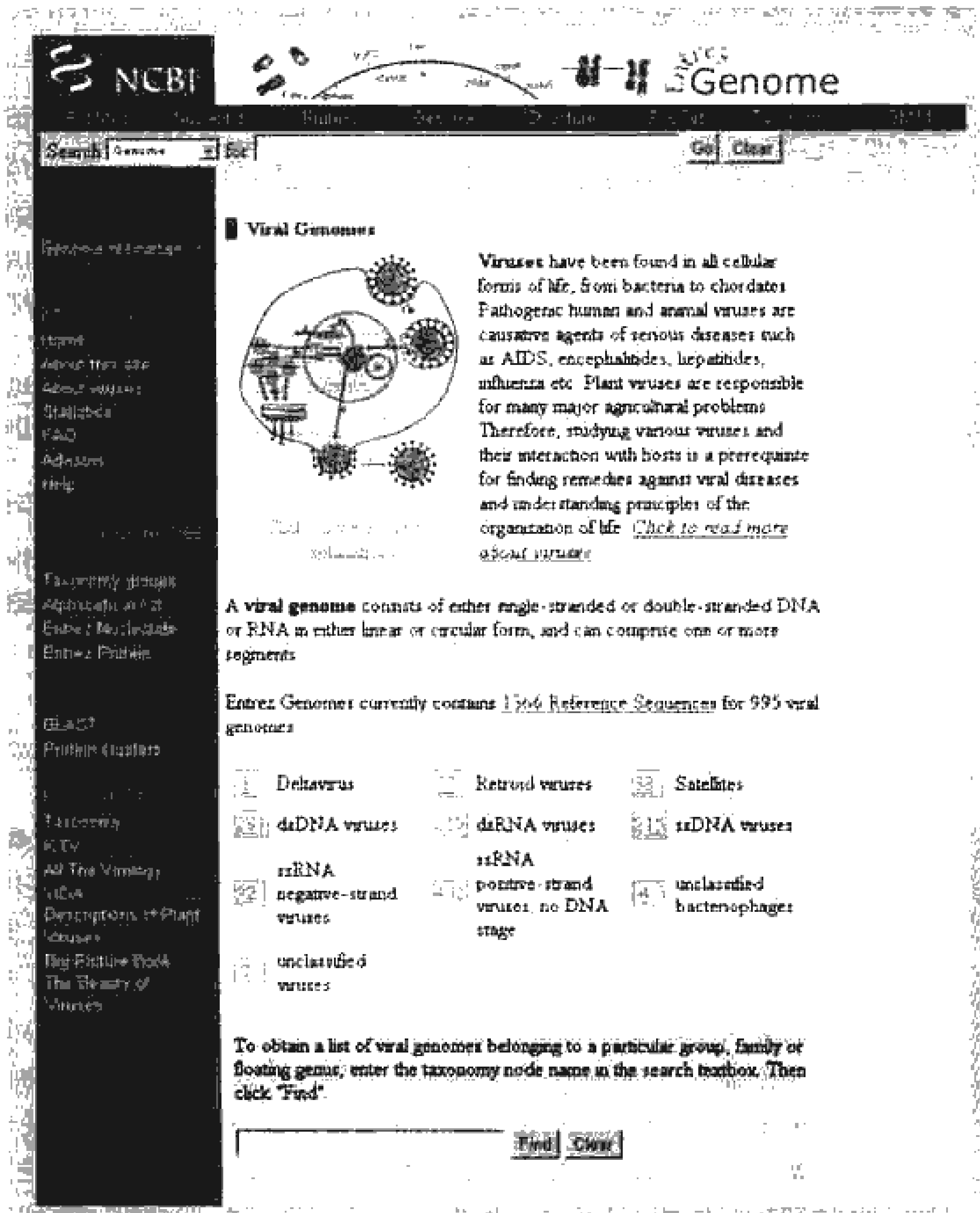


图 13-4 NCBI 的病毒基因组网页上提供了研究病毒所需的信息和资源

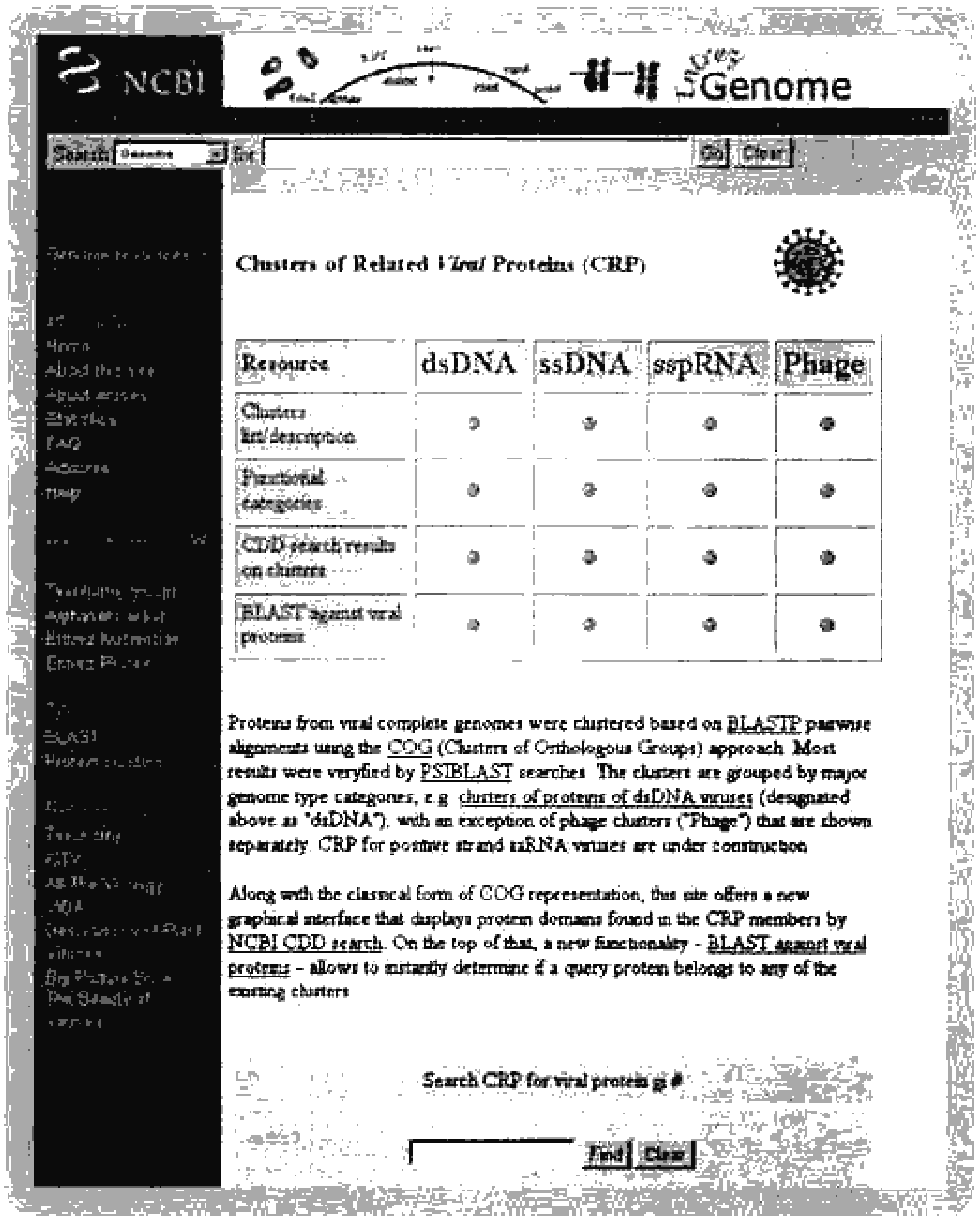


图 13-5 NCBI CRP 网站

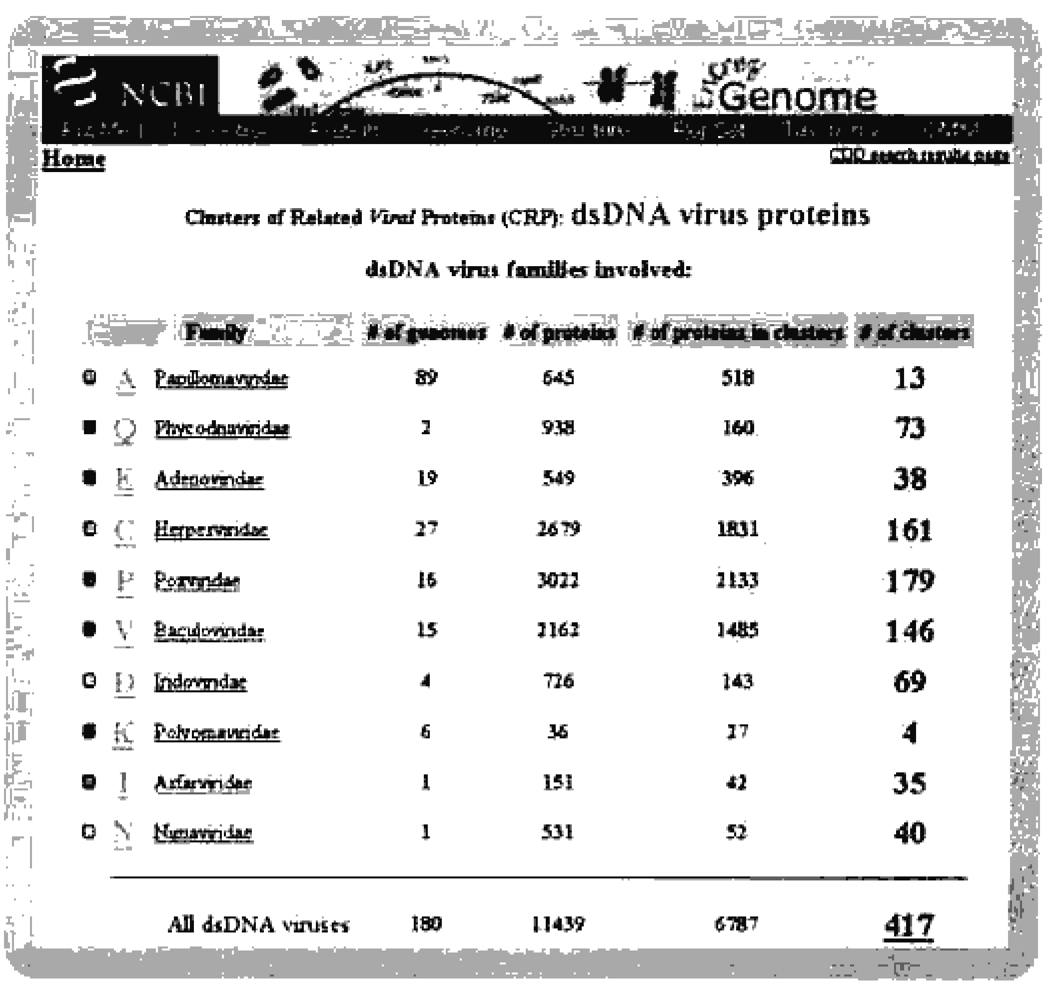


图 13-6 从 NCBI CRP 网页上能获得的双链 DNA 病毒家族的列表（其中包括疱疹病毒）

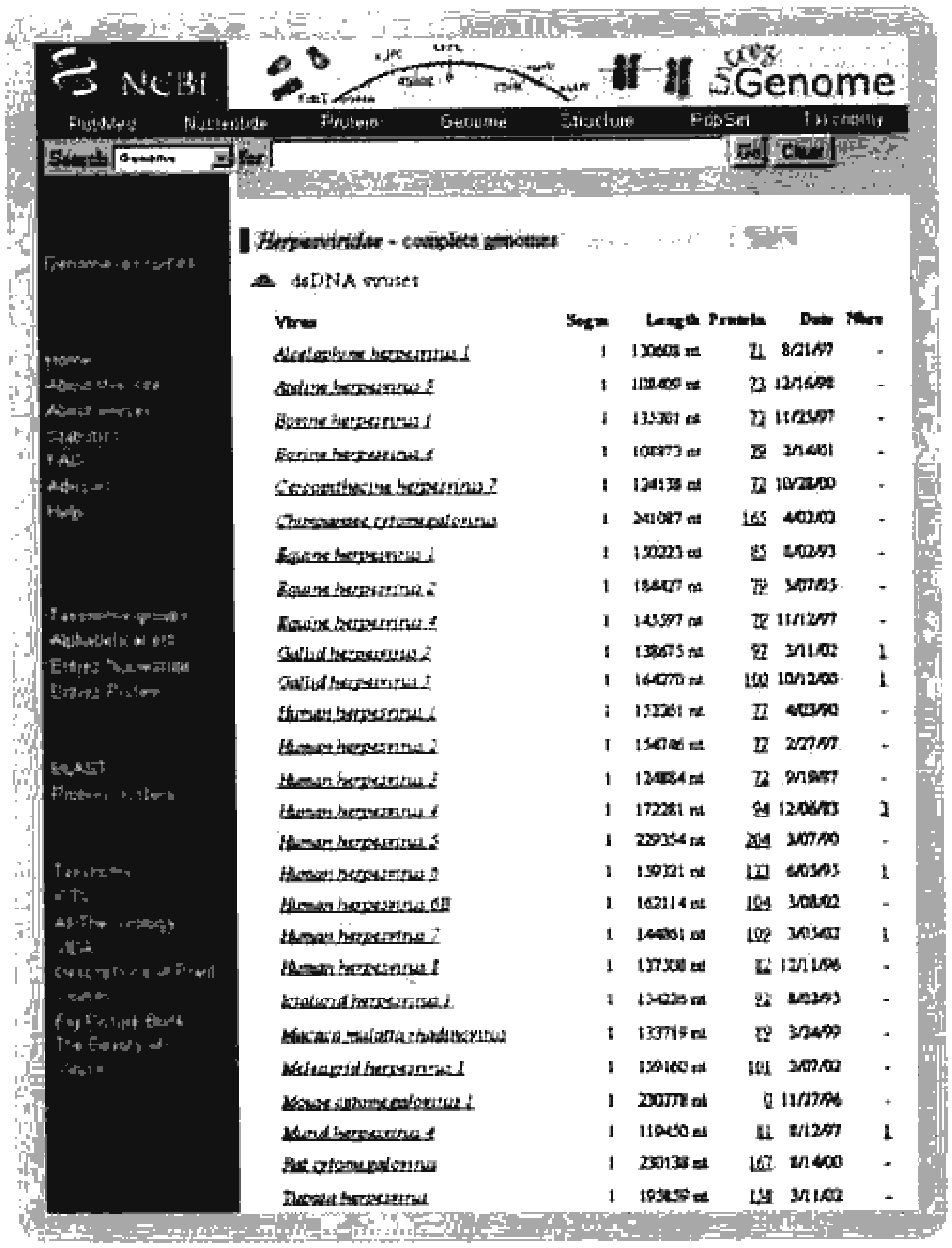


图 13-7 NCBI 疱疹病毒页面上提供了几十个疱疹病毒基因组的链接

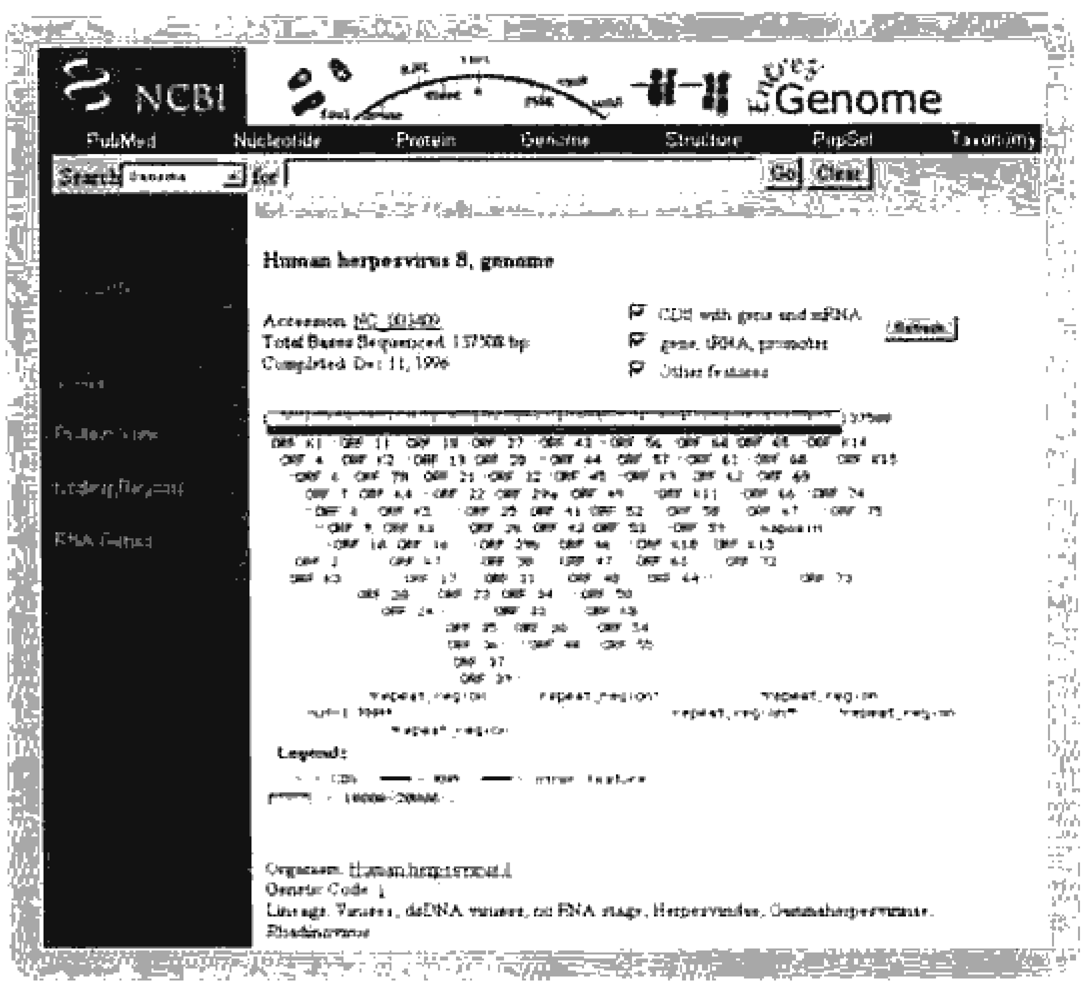


图 13-8 NCBI 上 HHV-8 基因组网页

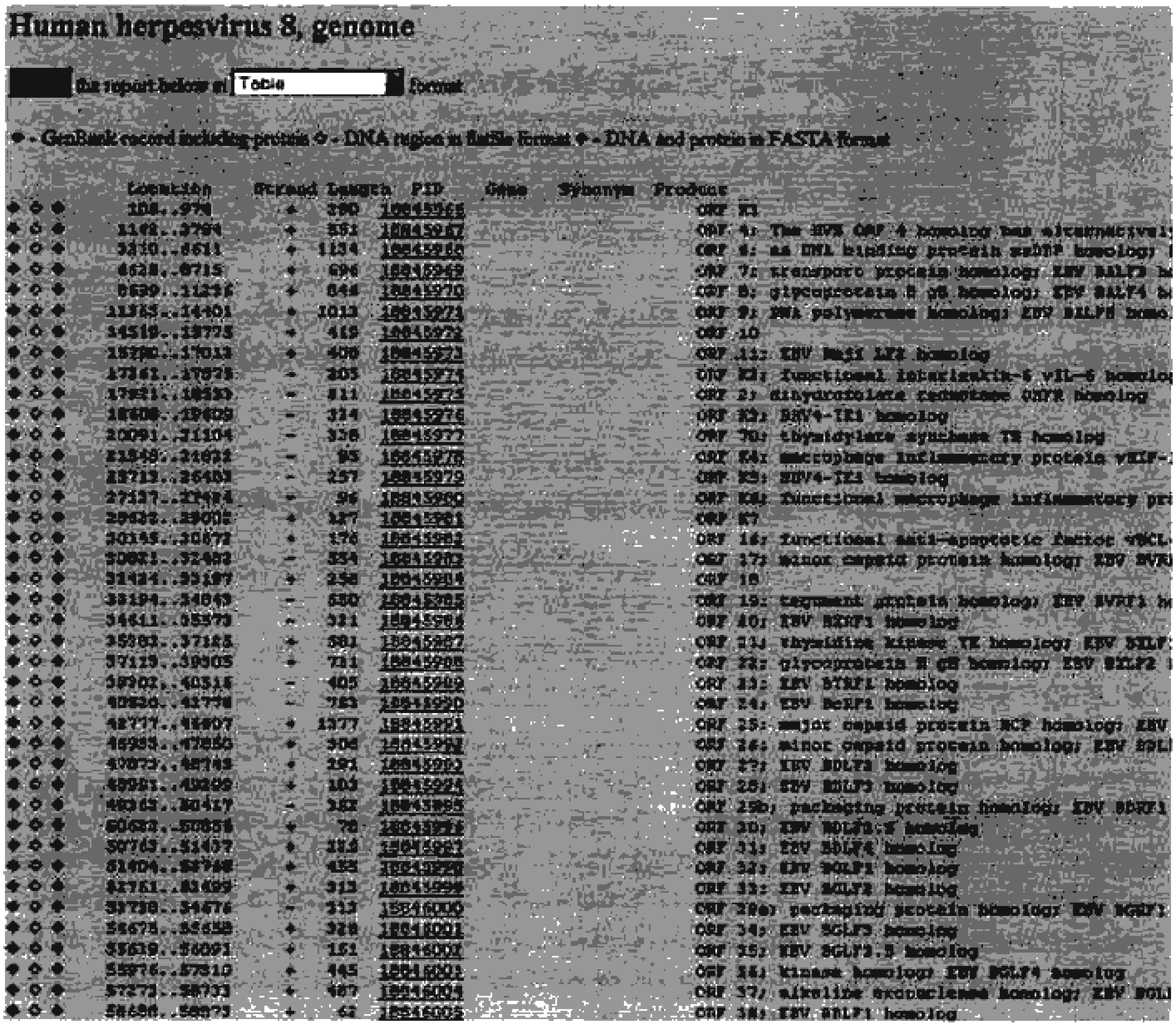


图 13-9 HHV-8 基因组页面上提供了这张含有 82 个开放阅读框的列表的链接

HHV-8 蛋白质包含病毒体结构和新陈代谢所需的蛋白质。有趣的是，它也包含多种与人类宿主蛋白质同源的蛋白，例如补体结合蛋白、细胞凋亡抑制剂 Bcl-2、二氢叶酸还原酶、干扰素调控蛋白、白介素 8（IL-8）受体、类神经细胞粘连分子黏附素和 D 类型细胞周期蛋白。

病毒的基因组如何从宿主生物体获得模体或整个基因？这能借由多种机制发生，包括重组、转座、剪接、移位和反向（McClure，2000）。例如，IL-8 受体基因是一个在真核细胞的生长和存活中发挥功能的基因。这个受体是 G 蛋白偶联受体大家族的一员，这个蛋白家族包括视紫红质，β-肾上腺素受体以及多种神经递质受体。包括 HHV-8 在内的多种病毒含有编码病毒 IL-8 受体的基因。以这个受体蛋白的序列作为查询序列（访问号 AAD46503）进行 blastp 搜索，揭示出一些病毒具有互不相同但是都与哺乳动物的 IL-8 受体紧密关联的蛋白（图 13-10）。据推测，当病毒感染一个哺乳动物细胞的时候，这种病毒 IL-8 蛋白质被表达从而有利于病毒的生存和生长（Wakeling 等，2001）。

两种互补的方法已经被用于进一步研究病毒基因（例如病毒 IL-8 受体）的功能和 HHV-8 侵染的机制。Paulose-Murphy 等（2001）合成了一个表示 88 个 HHV-8 的开放阅读框的微阵列，并使用这个微阵列测定了 HHV-8 病毒在人类细胞中处于裂解复制阶段时激活的基因的转录水平。他们测量了在诱导裂解型感染后一段时间内基因的表达水平，并描述了那些共表达的基因簇。这些基因可能在功能上相关。在较早阶段共表达的基因簇中的部分基因被猜测可能参

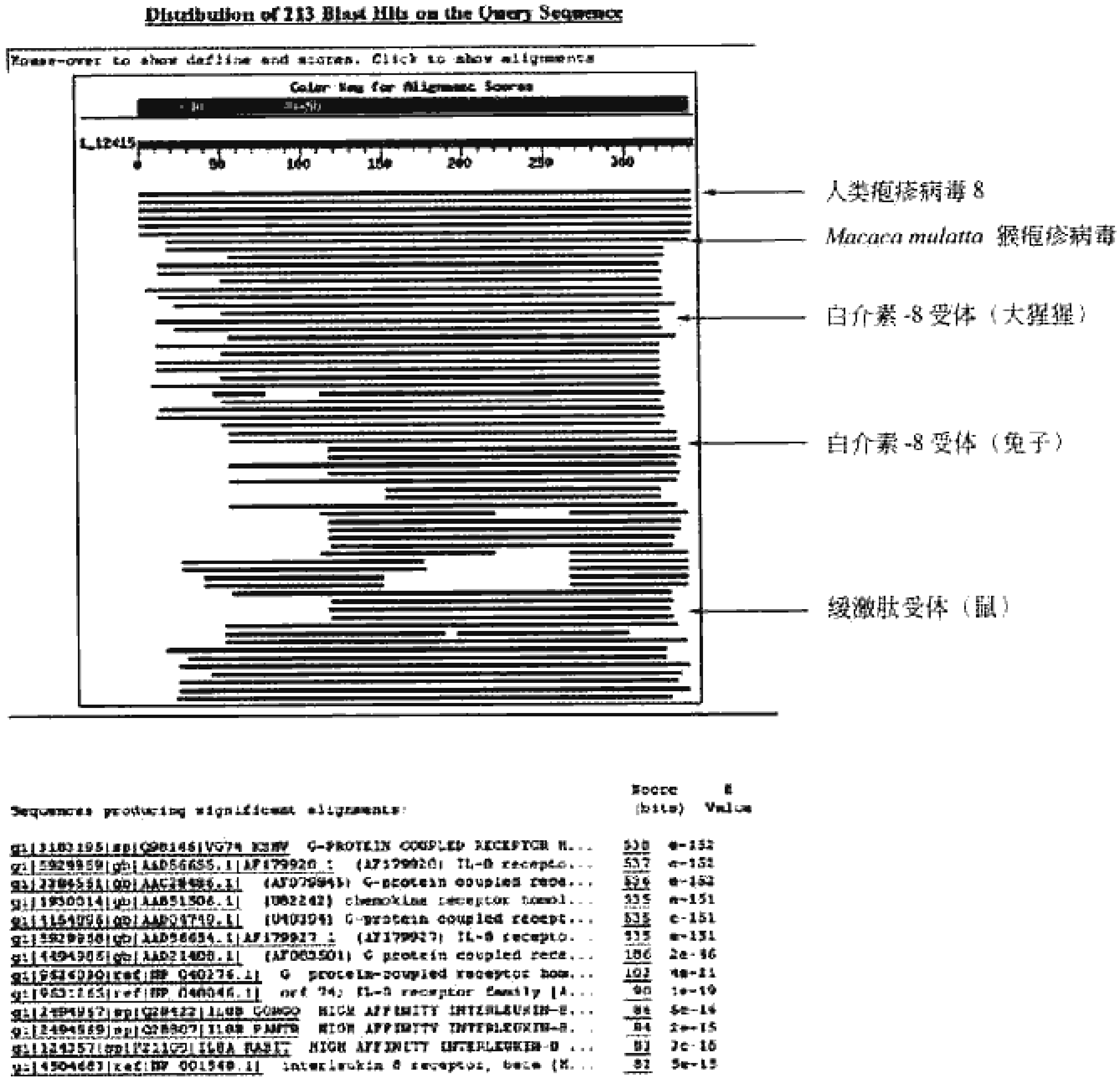


图 13-10 一个与哺乳动物 G 蛋白偶联受体蛋白超家族同源的病毒蛋白，这个蛋白在多种病毒中都有表达，包括卡波西瘤关联疱疹病毒（也称为 HSV-8）和鼠 γ-疱疹病毒、编码这个蛋白的基因很可能来源于哺乳动物，而后整合到了某些病毒的基因组中。在病毒侵染中，这种蛋白能够促进被侵染细胞的存活和生长

与了激活裂解循环的过程；另一部分则在病毒颗粒装配过程中负责编码蛋白质外壳（图 13-11）。在裂解循环中，与人类蛋白同源的病毒蛋白的表达自始至终都在进行。

在一项独立研究中，Poole 等（2002）用 HHV-8 侵染人类真皮中毛细血管的内皮细胞，测量了宿主细胞分别对潜伏性和裂解性病毒侵染的转录响应。HHV-8 的感染使内皮细胞从圆石状变成特征性的长梭形。卡波西瘤与许多其他病理症状有关，包括毛细血管形成和免疫调节病变。被 HHV-8 感染并调控的内皮细胞基因包括那些在免疫中发挥作用的干扰素应答基因，以及细胞凋亡、血管形成和细胞骨架相关蛋白的编码基因。这些研究有益于对宿主细胞对病毒侵染的响应进行准确的解释。

细胞凋亡是一种程序性细胞死亡，通过这种机制，宿主致摔被病毒感染的细胞，从而避免病原体在机体内广泛散布。然而，病毒也在进化中获得了操纵细胞凋亡通路的能力。同时，由于恶性感染病毒和肿瘤生长需要充足的血液供应，因而病毒有时能促进血管发育和生长。

艾滋病相关信息可在 NIH 网站（http://www.niaid.nih.gov/factsheets/aidsstat.htm）上获得；aids了解艾滋病流行现状信息可访问美国疾病和预防控制中心网站（http://63.1126.3.84/2002/Abstract/13996.htm），以及联合国艾滋病组织和世界卫生组织网站（http://www.unaids.org）。

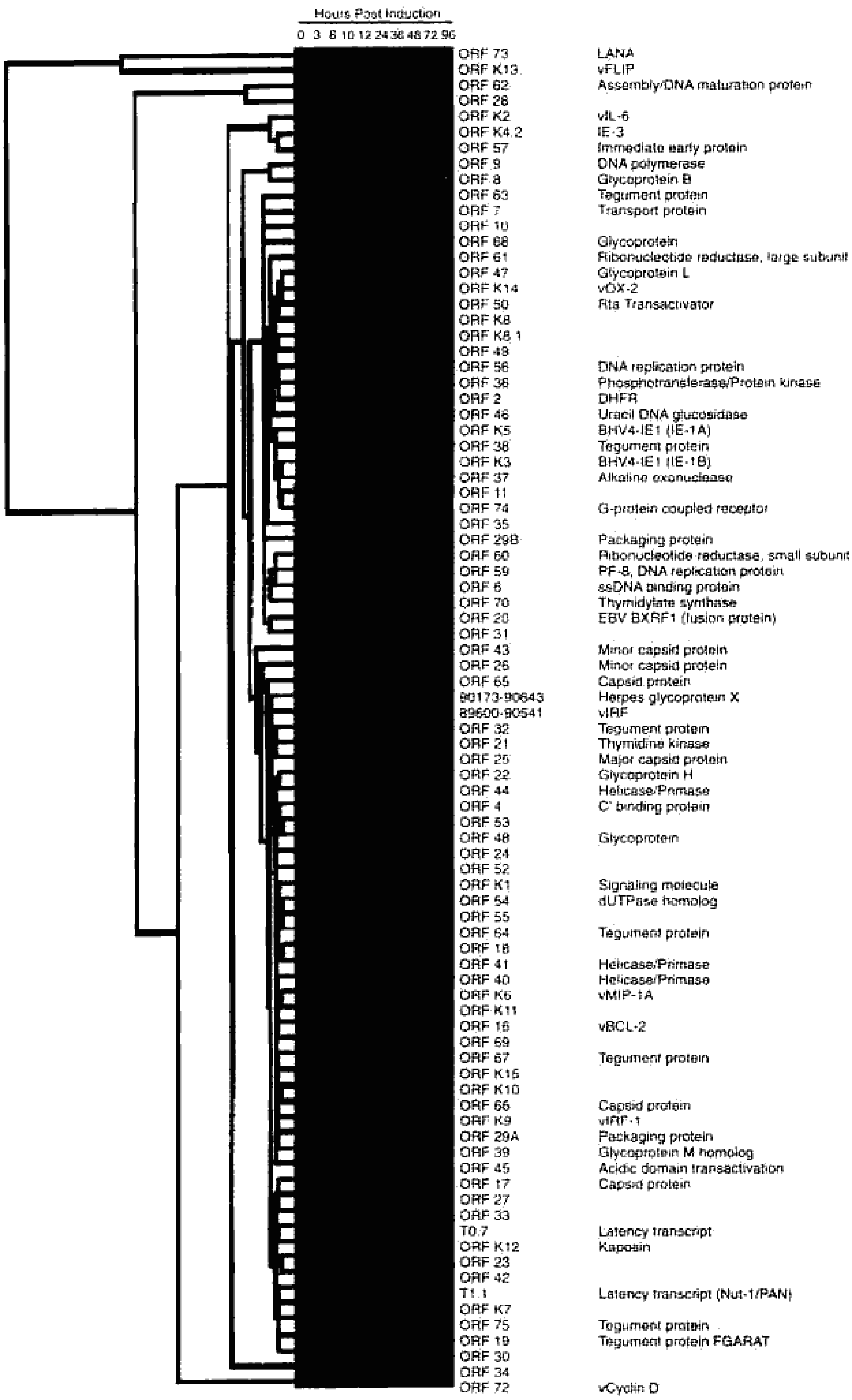


图 13-11 对 HHV-8 侵染的人类细胞中 HHV-8 病毒基因表达量微阵列测量数据进行双向层级聚类的结果。经过成对比较，表达速率相近的基因被分在相同的组。图中的列代表时间点，每一行代表一个开放阅读框的表达谱。经过归一化的基因表达速率用不同的颜色表示，以区别表达上调或下调。左侧的树状图是对表达谱相近的基因进行分组聚类的结果。这种方法有助于分辨在病毒侵染中发挥作用的基因。图片来自 Paulose-Murphy 等（2001），经授权使用

13.3.3 运用生物信息学方法研究人类免疫缺陷病毒

人类免疫缺陷病毒是导致艾滋病的病因（Meissner 和 Coffin，1999）。直到现在，艾滋病病毒还是绝对致命的。艾滋病的大部分症状不是由病毒直接引起的，而是由于病毒破坏了宿主的免疫系统而导致的。因此，艾滋病毒侵染使得其他病源生物乘虚而入从而导致疾病。

截至 2002 年底，全球已经有 4200 万人感染艾滋病，其中已经有 1600 万人死亡。仅 2002 年就新增大约 500 万感染者，其感染人数正在以大约每年 3%的速度增长。

艾滋病 1 型病毒（HIV-1）和艾滋病 2 型病毒（HIV-2）是属于慢病毒亚科的逆转录病毒。病毒可能最早来源于撒哈拉以南非洲，在那里艾滋病病毒的株型最多和感染率也最高（Sharp 等，2001）。在以 pol 蛋白全序列为基础建立的系统发生树中，灵长目慢病毒亚科分布在 5 个主要的谱系中［图 13-12(a)，箭头 1～箭头 5］（Hahn 等，2000）。这 5 个谱系是：

① 来自黑猩猩 *Pan troglodytes* 的猿免疫缺陷病毒（SIVcpz）以及 HIV-1；

② 来自黑色非洲白眉猴 *Cerecocebus atys* 的猿免疫缺陷病毒（SIVsm）以及 HIV-2 和短尾猿猿免疫缺陷病毒（SIVmac）；

③ 来自非洲绿猴的猿免疫缺陷病毒（SIVagm）；

④ 来自赛克斯猴 *Cercopithecus albogularis* 的猿免疫缺陷病毒（SIVsyk）；

⑤ 来自弯钩长尾猴 *Cercopithecus lhoesti* 的猿免疫缺陷病毒（SIVlhoest）、suntailed 猴 *Cercopithecus solatus* 的猿免疫缺陷病毒（SIVsun），以及来自西非洲产大狒狒 *Mandrillus sphinx* 的猿免疫缺陷病毒（SIVmnd）。

某种疾病的流行率指的是在某一时刻，患有该疾病的人群占总人数的比例。流行率并不描述某个体在何时感染上该疾病。发病率指的是在一段时间内新增疾病感染者的频率。例如，我们通常用类似"一年里每 1000 人中新增 10 个病例"的方式描述发病率。

如图 13-12(a) 所示，系统发育分析得出的一个显著特征是病毒以一种依赖其宿主的方式进化（Hahn 等，2000）。与来自其他物种的病毒相比，感染除人类外其他灵长目动物的病毒之间具有更多的相关性。例如 HIV-2 可能来源于黑色非洲白眉猴的传播，支持这一观点的证据有 5 个（Hahn 等，2000）：

① HIV-2 和 SIVsm 的基因组结构类似；

② HIV-2 和 SIVsm 的系统发生相关（见图 13-12，箭头 4）；

③ SIVsm 在其天然宿主中广泛传播；

④ HIV-2 感染者和 SIVsm 天然宿主所在地域相同；

⑤ 存在可能的传染途径，例如在市场上人可能接触到黑猩猩的血液。

相似的观点也适用于 HIV-1。这种病毒很可能是在 1930～1940 年间由 SIVcpz 在非洲黑猩猩和人之间交叉传染产生的。HIV-1 有三种主要类型，分别是 M、N 和 O，这与 SIVcpz 的三种传播类型相一致。M 是 HIV-1 的主要类型；O 和 N 分别是一种与 M 亲缘关系较远

截至 2003 年 6 月，GenBank 中储存有 90000 条 HIV-1 的核酸记录。在 NCBI 的物种分类浏览器中输入"HIV-1"就能访问到这些序列。如果限制搜索条件，只查询 Entrez 核酸数据库中的 RefSeq，那么将只能看到一条搜索结果：HIV-1 的全基因组序列（NC_001802）。

Entrez 基因组部分（http://www.ncbi.nlm.nih.gov/entrez/query.fcgi? db=Genome）涵盖了数百种病毒。截至 2003 年 6 月，该网站上储存有 1300 个完整的病毒基因组序列。另外还有超过 100 个噬菌体的序列和 30 多个类病毒（具有感染性的能够导致某些植物疾病的 RNA 基因组）。

的类型。由 Env 蛋白全序列构建的系统发生树上可以清楚地看到 HIV-1 的这三个主要类型[图 13-12(b)，箭头 6～箭头 8]（Hahn 等，2000）。

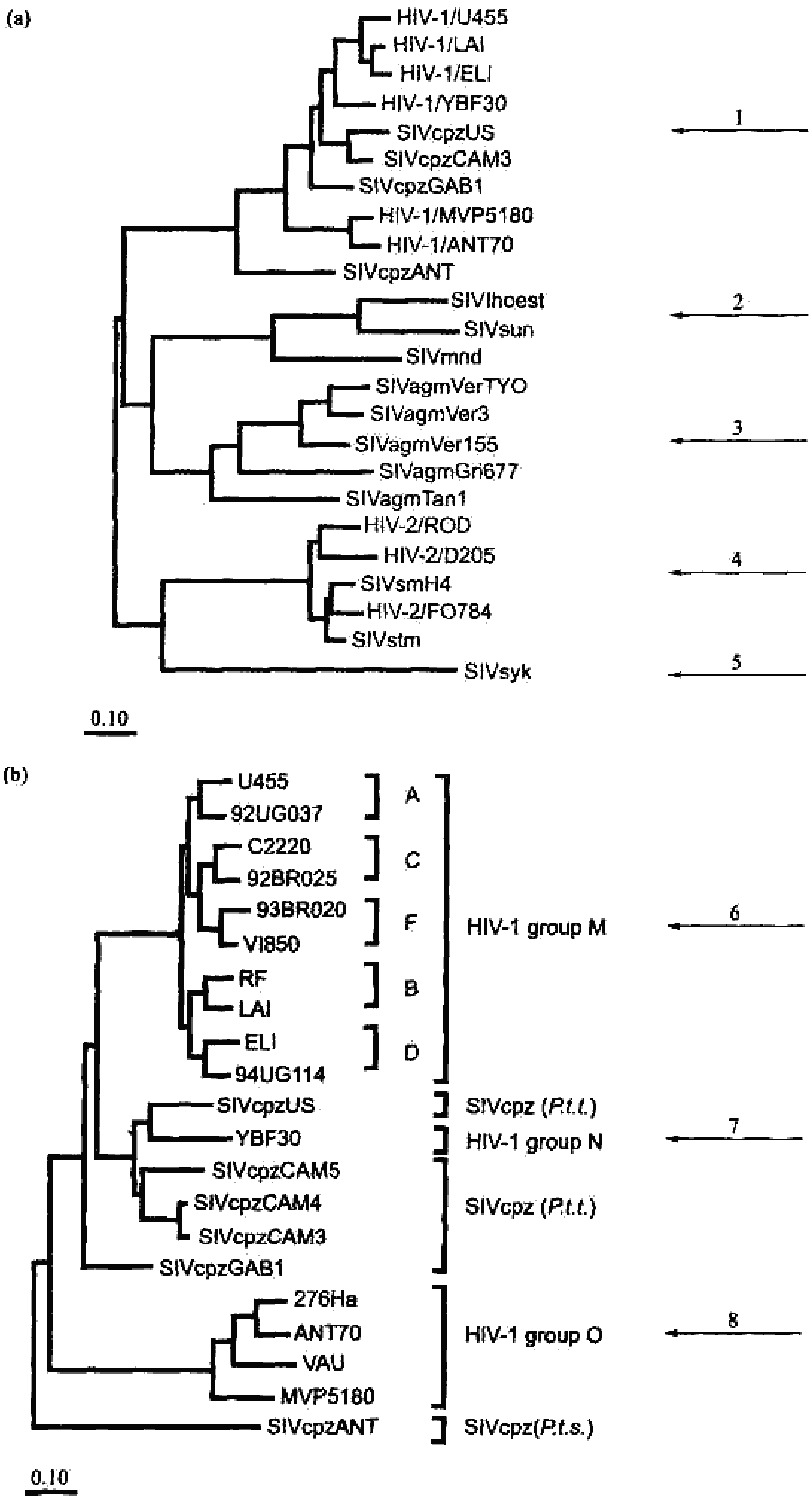


图 13-12 灵长类动物慢病毒亚科病毒的进化关系。(a) 对 Pol 蛋白全长序列进行多序列比对并采用最大相似性方法构建的进化树，其中有 5 个主要分支（箭头 1～箭头 5）。图下方的标度代表每个位点发生 0.1 次氨基酸替换所需的进化距离。(b) 对 HIV-1/SIVcpz 病毒 Env 蛋白序列进行多序列比对并采用最大相似性方法构建的进化树。HIV-1 的 3 个主要分支（M，N，O；箭头 6～箭头 8）在图中被清楚地区分开。图中所使用的标度与 (a) 相同。图片来自 Hahn 等（2000），经授权使用

在图 2-12 中我们已经看过了 HIV-1 在 NCBI 中的相应条目，它的长度为 9181 个碱基，编码 9 种蛋白质。尽管 HIV-1 的基因组很小并且只编码少量基因产物，但 GenBank 中将会很快存储超过 100000 条它的核酸序列记录和同样数量的蛋白质序列记录。这是由于 HIV-1

变异的速度极快，能够产生许多 M、N 和 O 亚类的变体。也因此，研究人员必须经常测定 HIV-1 的序列。病毒学家面临的一个主要挑战就是如何处理如此大量的数据以及如何从这些数据中找到治疗或治愈艾滋病的方法。接下来我们将介绍帮助研究艾滋病病毒序列数据的两个生物信息学数据库资源：NCBI 和 LANL。

在 Entrez 基因组部分，病毒分类下有一个病毒列表，列表中含有病毒名和 HIV-1 的编号（NC_001802），并按照字母序排列。点击病毒名将会链接到 NCBI 物种分类浏览器，使用者将会看到 HIV-1 的系统分类（病毒；逆转录病毒目；逆转录病毒科；慢病毒亚科；灵长类慢病毒属）以及 HIV-1 变体的链接。在 Entrez 基因组部分页面上，点击编号 NC_001802 将会链接到 Entrez 核酸（Genbank）数据库中 HIV-1 的条目。HIV-1 基因组是一个有 9181 个碱基的单链 RNA 分子。通过点击页面左栏的“编码区域”，使用者将会链接到病毒的编码区信息。通过这个方法可以很方便地查询对 HIV-1 基因组上感兴趣的 DNA（或蛋白质）的核苷酸（或氨基酸）序列。

13.3.4 运用生物信息学方法研究 HIV-1

13.3.4.1 使用 NCBI 数据库资源

NCBI 网站提供了多种帮助研究逆转录病毒（包括 HIV-1）的方法。与上文描述的研究 HHV-8 的方法一样，可以通过 NCBI 的 Entrez Genome 站点访问 HIV-1 的信息。

另外，NCBI 也提供专门用于研究逆转录病毒的资源（图 13-13）。这个站点包含了如下资源：

① 一个以 BLAST 搜索为基础的基因类型查询工具；

② 一个专门用于逆转录病毒的多序列比对工具；

③ 50 个逆转录基因组的参考序列；

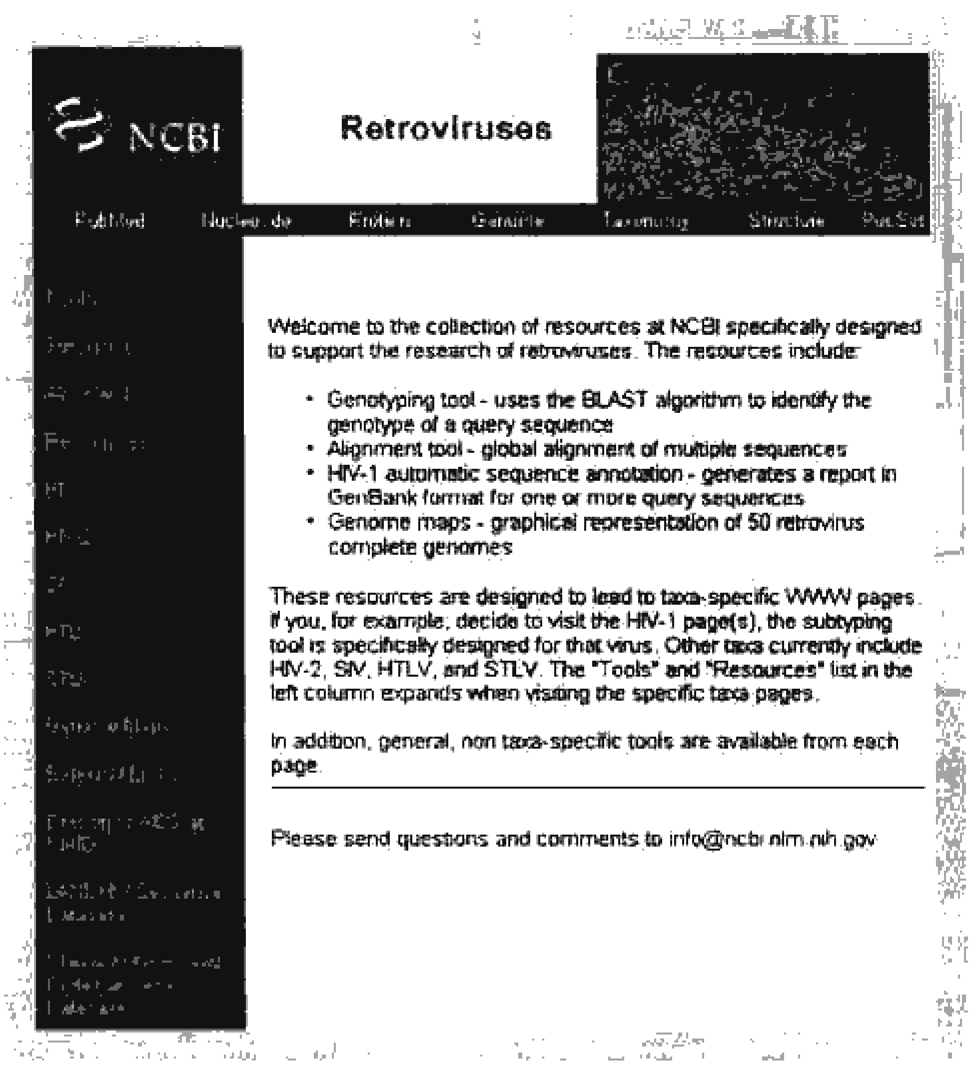


图 13-13 NCBI 提供的逆转录病毒数据资源（见 http://www.ncbi.nlm.nih.gov/retroviruses/）

④ 专门用于研究 HIV-1、HIV-2、SIV、人类 T 淋巴细胞 1 型病毒和 STLV 的工具页面；

⑤ 最近一周内发表的关于逆转录病毒的文献清单；

⑥ 最近一周内 GenBank 新增的逆转录病毒序列清单（GenBank 上每周都会增加数百条新的 HIV-1 序列）；

⑦ 到外部其他逆转录相关网站的链接。

在 NCBI 主页上（http://www.ncbi.nlm.nih.gov/）选择"逆转录病毒资源"（http://www.ncbi.nlm.nih.gov/retroviruses/）。

13.3.4.2 使用 LANL 数据库

美国洛斯阿拉莫斯国家实验室运行和维护了 4 个有关 HIV-1 的数据库及其网站。其中 HIV 序列数据库存储了大量全面的 HIV 序列数据。它允许用户通过名称、访问编号、PubMed 标识、取样地区以及易感染地区来搜索特定序列。用户可以获取包含多序列比对结果的序列信息，或者来自同一病人的不同 HIV 病毒序列。这个站点还包含了一系列专用工具，包括：

LANL HIV 数据库位于 http://hiv-web.lanl.gov/。这个站点提供 4 个数据库，分别是序列数据库、抗药性数据库、免疫数据库和疫苗试验数据库。在 HIV 序列数据库中可以找到 HIV 病毒亚型分布地图工具 http://hiv-web.lanl.gov/content/hiv-db/geography.comp。

① 一个 HIV BLAST 服务器；

② 一个用于计算同义和非同义碱基替换率的程序——SNAP（Synonymous/Non-synonymous Analysis Program）；

③ 一个用于识别可能由染色体重组产生的病毒镶嵌序列片段的程序——RIP（Recombinant Identification Program）；

④ 一个基于 HMMER 软件（第 10 章）的多序列比对工具——MPAlign（Gaschen 等，2001）；

⑤ 一个根据距离打分对序列进行主要组成分析（第 7 章）的软件——PCoord（Principal Coordinate Analysis）；

⑥ 一个标有各地 HIV 感染水平并能显示 HIV 各亚类分布的地图工具（既可显示世界范围，也可以只针对某一大陆）［图 13-14(a)，图 13-14(b)］。

LANL 蛋白酶突变与药物关系图的网址是 http://resdb.lanl.gov/Resist DB/protease mutation map.htm

LANL 还包含其他一些为运用生物信息学方法分析 HIV-1 和相关病毒提供重要工具的数据库。HIV 抗药性数据库允许使用者浏览 *HIV-1* 基因，并标示出哪些药物的药性会受到基因产物氨基酸替换的影响（图 13-15）。HIV 抗药性数据库还提供对这些信息的直观的图形显示，图上每个可点击的点均对应一个 HIV 蛋白的完整序列［如蛋白酶：图 13-16(a)，箭头 1］，这些点的纵坐标代表不同的药物。图上还标出了能够导致 HIV 对某类药物产生抗性的氨基酸替换，点击这些替换的氨基酸［图 13-16 (a)，箭头 2］可以进一步看到发生氨基酸替换的突变细节［图 13-16(b)］。

美国国家癌症研究院下属的 Frederick 癌症研究和开发中心建立的 HIV 蛋白酶数据库（http://mcll.ncifcrf.gov/hivdb/index.html）包含有数百条 HIV-1、HIV-2 和 SIV 的蛋白酶结构信息。

13.3.5 针对麻疹病毒的生物信息学分析

在人类历史上，麻疹是最为致命的疾病之一。今天，在许多国家，它仍是幼儿生命的首要威胁，每年有超过 100 万的婴儿死于麻疹（Johnson 等，2000）。尽管麻疹疫苗的使用有

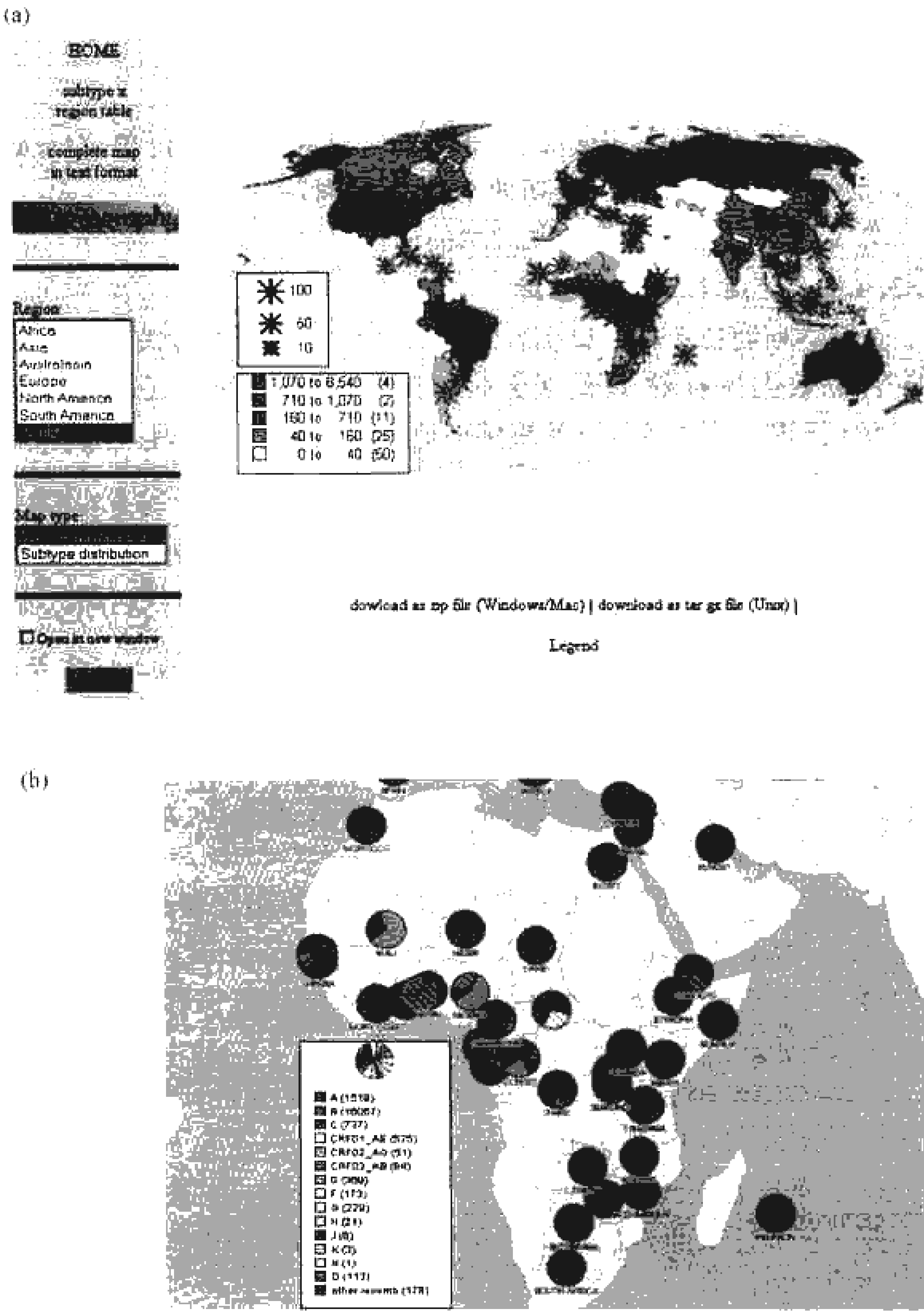
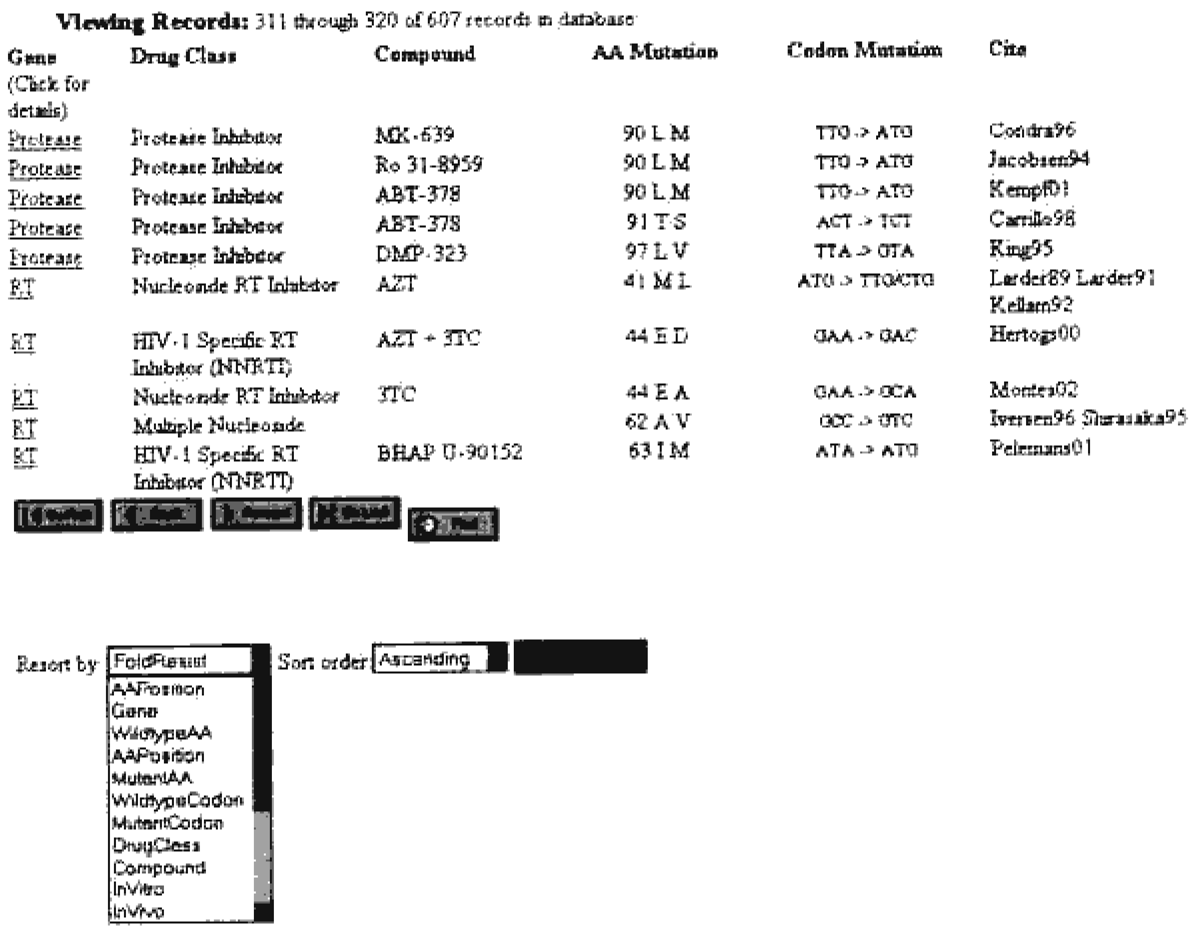


图 13-14 LANL 提供的地图显示工具可以让使用者观察 HIV 各亚型在（a）世界范围或（b）某大洲范围（图中所示为非洲）的感染情况。在（a）中显示了各地区感染的主要 HIV 亚型，不同颜色的正方形表示在 LANL 数据库中存储的各国 HIV 的序列条数，不同大小的星号代表非主要亚型在全部感染人群中所占的比例（星号越大代表 HIV 的多样性越高）。各地区 HIV 各亚型的感染比例在（b）中用饼状图表示。这个地图显示工具的网址是 http://hiv-web.lanl.gov/content/hiv-db/geography/map right.html

在麻疹病毒疫苗投入使用前，美国每年约有 4500000 例麻疹病人（约 450 人死于麻疹）。见 http://www.cdc.gov/nip/diseases/measles/q&a.htm。

效降低了麻疹的发病率和死亡率，但是在婴儿出生后 9 个月之前，由于婴儿自身免疫系统尚不成熟以及母体抗体的存在，疫苗难以接种成功。麻疹病毒通过飞沫传播感染呼吸道内皮细胞。

麻疹病毒是一种属于副黏病毒科的疹病毒，腮腺炎病毒和呼吸道合胞体病毒也属于这一科。麻疹病毒的全基因组序列可以通过 NCBI Entrez 基因组资源库访问到（编号 NC _ 001498）。麻疹病毒由一个不分段的反义 RNA 基因组以及包被它的衣壳体构成。它的基因组长 15894bp，编码 7 种蛋白质。这些编码序列可以通过点击 Entrez 条目记录左栏的“编码区”选项查看［图 13-17(a)］。6 个基因分别记作 *N* 基因（核壳体基因），*P* 基因（磷蛋白基因），*M* 基因（基质基因），*F* 基因（融合基因），*H* 基因（红细胞凝集素基因）和 *L* 基因（大聚合酶基因）［图 13-17(b)］。有人预测 *P* 基因可能通过改变转录起始点和开放阅读框编码另



Viewing Records: 311 through 320 of 607 records in database

Gene (Click for details)	Drug Class	Compound	AA Mutation	Codon Mutation	Cite
Protease	Protease Inhibitor	MK-639	90 L M	TTG -> ATG	Condra96
Protease	Protease Inhibitor	Ro 31-8959	90 L M	TTG -> ATG	Jacobsen94
Protease	Protease Inhibitor	ABT-378	90 L M	TTG -> ATG	Kempf01
Protease	Protease Inhibitor	ABT-378	91 T S	ACT -> TCT	Carrillo98
Protease	Protease Inhibitor	DMP-323	97 L V	TTA -> GTA	King95
RT	Nucleoside RT Inhibitor	AZT	41 M L	ATG -> TTG/CTG	Larder89 Larder91 Kellam92
RT	HIV-1 Specific RT Inhibitor (NNRTI)	AZT + 3TC	44 E D	GAA -> GAC	Hertogs00
RT	Nucleoside RT Inhibitor	3TC	44 E A	GAA -> GCA	Montes02
RT	Multiple Nucleoside		62 A V	GCC -> GTC	Iversen96 Shirasaka95
RT	HIV-1 Specific RT Inhibitor (NNRTI)	BHAP U-90152	63 I M	ATA -> ATG	Pelemans01

Resort by FoldResult Sort order Ascending

AAPosition
Gene
WildtypeAA
AAPosition
MutantAA
WildtypeCodon
MutantCodon
DrugClass
Compound
InVitro
InVivo

图 13-15 HIV 药物抗性数据库汇集了导致抗 HIV 药物抗性的 HIV 基因的突变。其提供的浏览工具（网址：http://resdb.lanl.gov/ResistDB/default.html）允许使用者搜索所有的可发生此类突变的 *HIV-1* 基因，并且这些数据可以依照多种标准进行排序

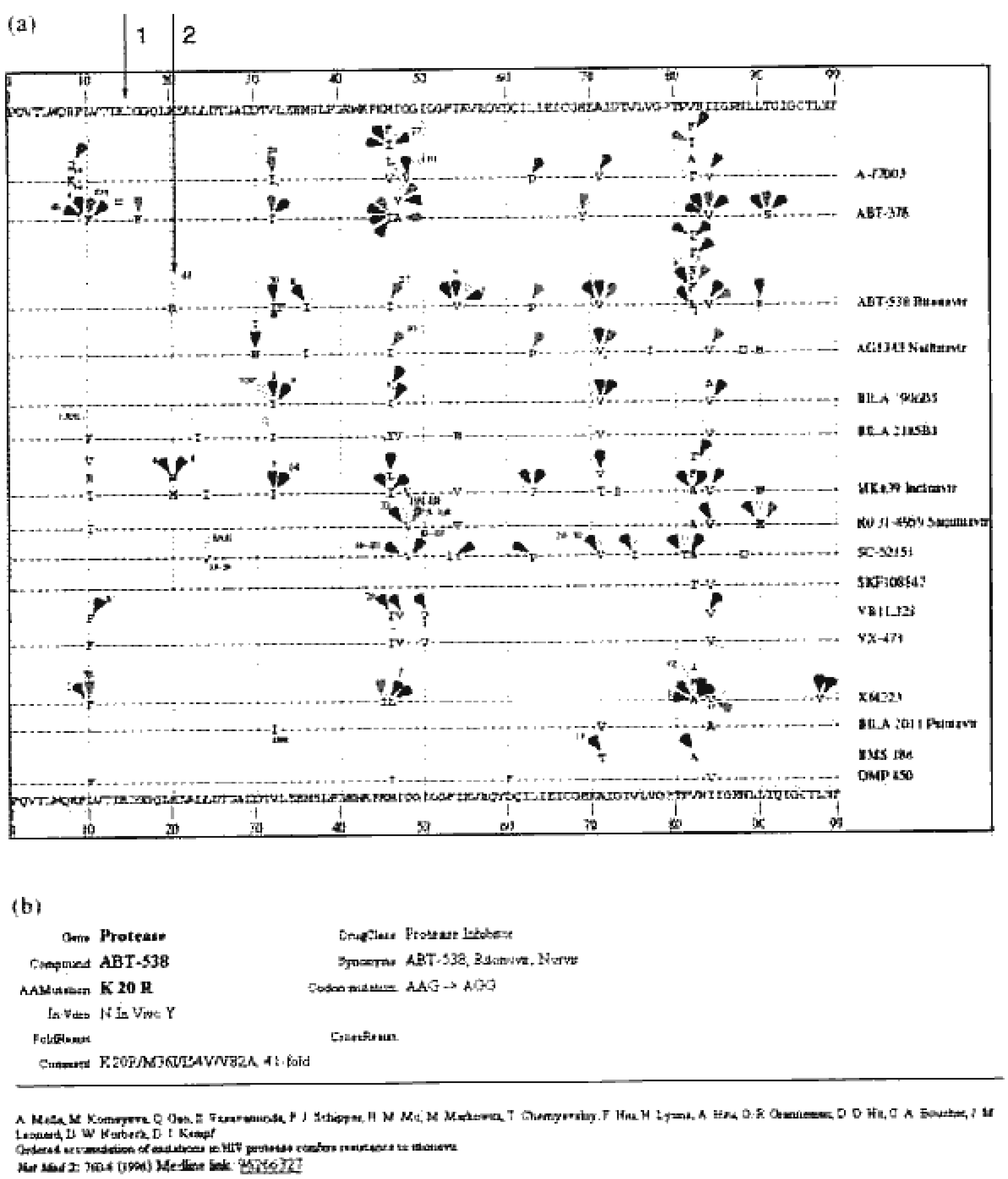
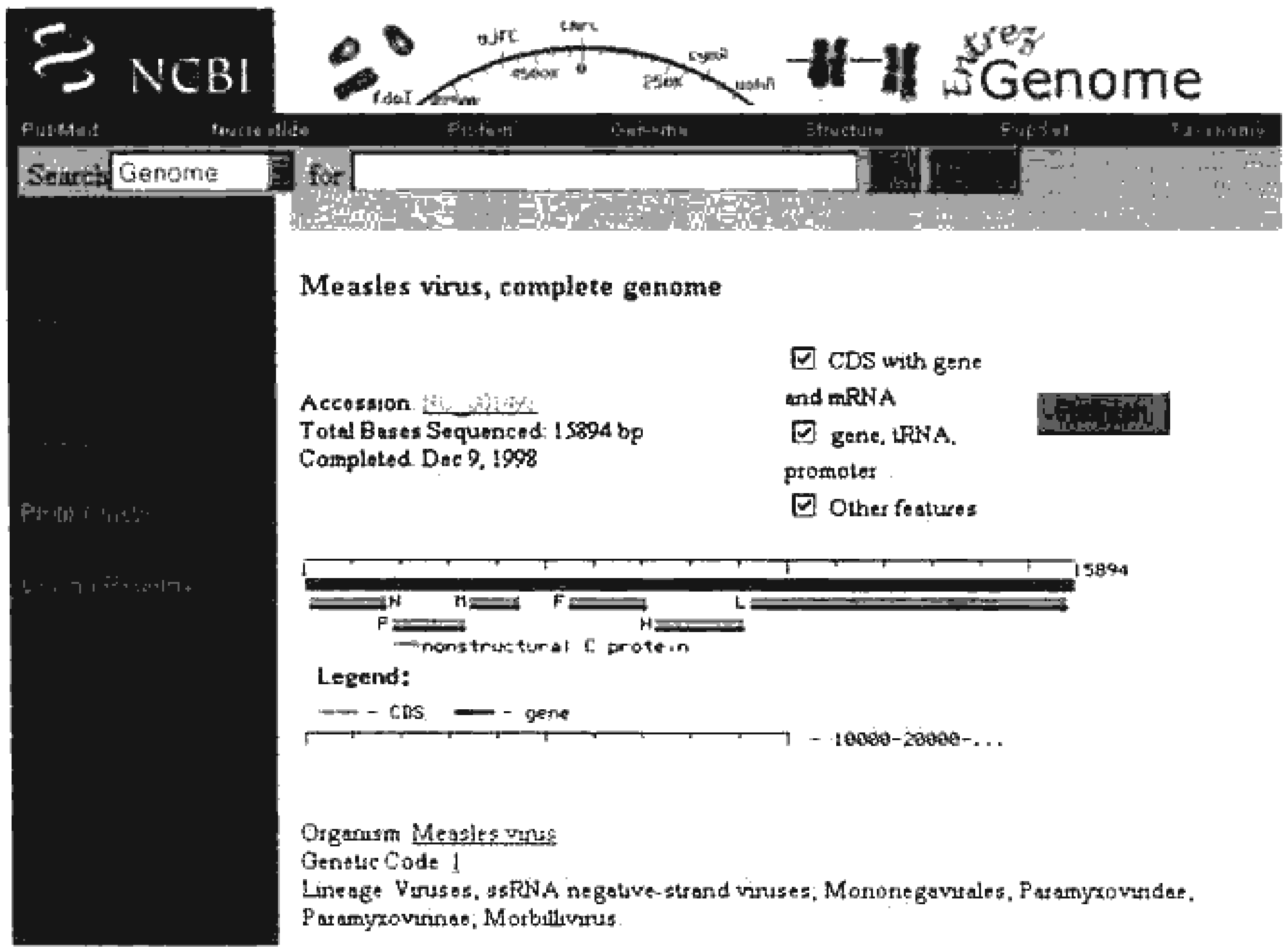
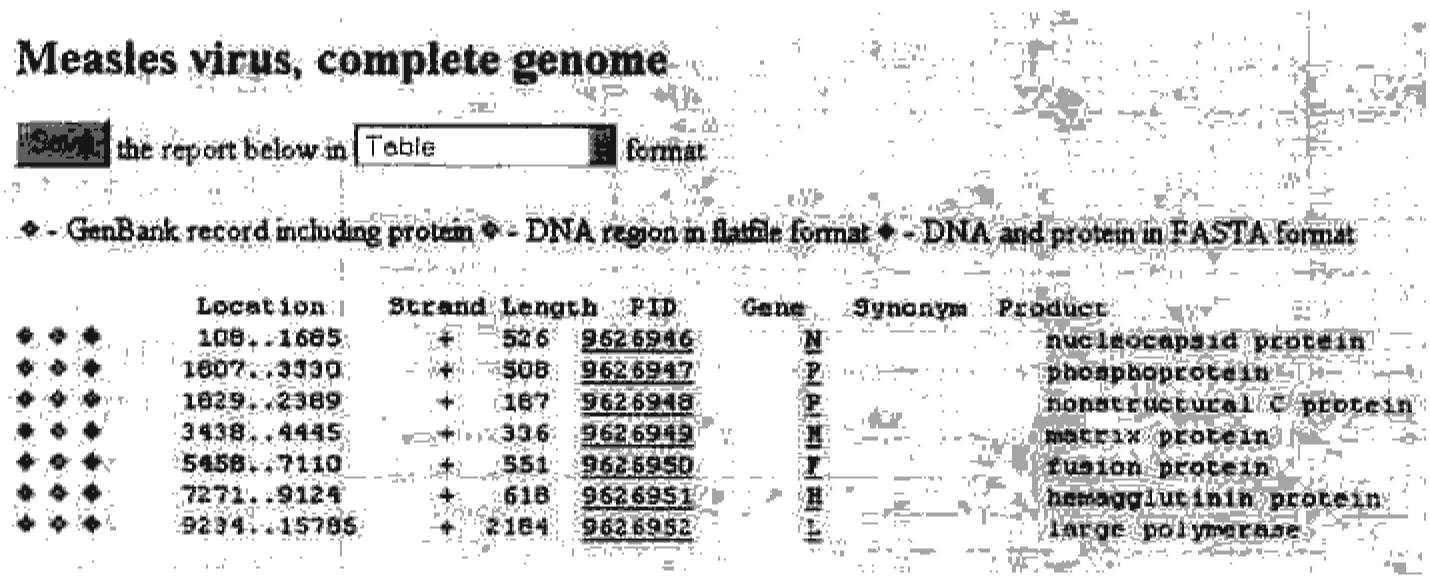


图 13-16 (a) LANL 网站提供的 HIV-1 蛋白突变对不同药物影响的对应图。图中每一行代表一种药物（见右侧标记）。野生型 HIV 蛋白（HXB2 株）的序列列在图的顶部和底部（箭头 1），虚线代表野生型蛋白氨基酸序列的位置，可能导致药物抗性的突变位置都被标出。箭头 2 所指的是一个 K→R（赖氨酸→精氨酸）氨基酸突变。箭头旁的小数字（41）代表这个突变引起的抗性提高倍数。那些有彩色小箭头所指的突变代表它们是协同组合突变的一部分。(b) 通过点击突变的位点（箭头 2 所指处），将看到该突变所带来的影响的细节

(a)



(b)



Measles virus, complete genome

the report below in Table format

◆ - GenBank record including protein ◆ - DNA region in flatfile format ◆ - DNA and protein in FASTA format

	Location	Strand	Length	PID	Gene	Synonym	Product
◆ ◆ ◆	108..1685	+	526	9626946	N		nucleocapsid protein
◆ ◆ ◆	1807..3330	+	508	9626947	P		phosphoprotein
◆ ◆ ◆	1829..2389	+	187	9626948	P		nonstructural C protein
◆ ◆ ◆	3438..4445	+	336	9626949	M		matrix protein
◆ ◆ ◆	5458..7110	+	551	9626950	F		fusion protein
◆ ◆ ◆	7271..9124	+	618	9626951	H		hemagglutinin protein
◆ ◆ ◆	9234..15785	+	2184	9626952	L		large polymerase

图 13-17 对麻疹病毒的基因组分析。(a) 通过 NCBI Entrez 网站访问麻疹病毒条目。通过点击左栏上"编码区域"选项，将看到页面 (b) 蛋白编码基因列表。这些基因分别用字母 *N*（核壳体基因）、*P*（磷蛋白基因）、*M*（基质基因）、*F*（融合基因）、*H*（红细胞凝集素基因）和 *L*（大聚合酶基因）表示。经预测，*P* 基因可能通过改变转录起始位点和阅读框编码另一个蛋白（非结构蛋白 C）

另一种副黏病毒科病毒是古老的牛疫病病毒（Barrett 和 Rossiter，1999）。这种病毒对人类和反刍动物都有极大的危害。

一种蛋白（非结构蛋白 C）。通过点击 Entrez 核酸记录，选择"图像"显示选项可以很方便直观地观察到上述可能性（图 13-18）。图像显示出麻疹病毒基因组编码非结构蛋白 C 的位置。

这 6 个蛋白的功能已经确认，N 结合并环绕基因组 RNA，P 与 L 形成复合体参与 RNA 合成，H 和 F 插入病毒外膜成为病毒表面蛋白，其中 H 与宿主细胞膜外受体结合介导病毒进入细胞，F 是

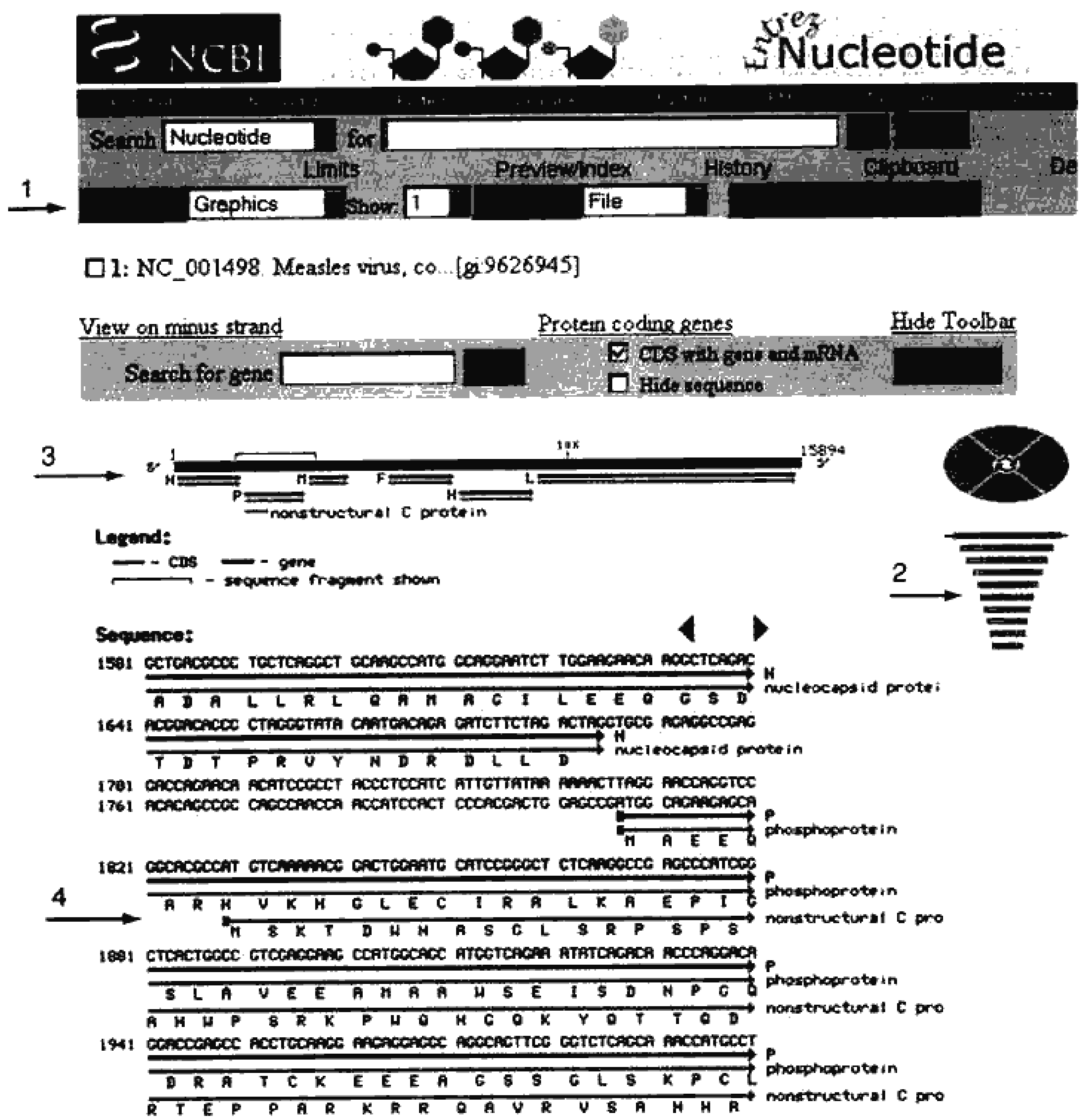


图 13-18　对 DNA 序列及其编码的蛋白序列的并列图形化显示（箭头 1）。可以很方便地对图形进行缩放以便于观察序列的整体或细节（箭头 2）。箭头 3 所指为一段麻疹病毒基因组及其编码的蛋白序列的总体情况。从下方的细节显示中，我们可以看到基因 *P* 通过采用不同的起始位点和阅读框编码 2 种不同的蛋白（箭头 4）

一种融合蛋白，促使病毒穿过宿主细胞质膜，M 在病毒核糖核蛋白 N、P、L 与包被蛋白 H、F 之间起连接作用。这些蛋白的功能都能通过 BLAST 工具搜索到。例如非结构蛋白 C，通过对去冗余（nr）数据库进行 blastp 搜索，发现这种蛋白与牛疫病毒、犬热病病毒以及海豚麻疹病毒编码的蛋白具有同源性。用病毒的红细胞凝集素对去冗余数据库进行 blastp 搜索，发现这种蛋白属于 Pfam 蛋白质家族中的一员（pfam00423，红细胞凝集素-唾液酸苷酶家族），而且有数百个与麻疹病毒红细胞凝集素相匹配的条目。限制关键词为"hemoggluti-nin Not measles virus [Organism]"重新搜索，搜索结果减少为 30 多个同源病毒红细胞凝集素的条目。使用 PSI-BLAST 工具还可以搜索到其他病毒的红细胞凝集素的条目，例如人类副流感病毒、腮腺炎病毒以及火鸡鼻气管炎病毒。

13.3.6 其他生物信息学资源

Viral Genome Organizer 是由加拿大维多利亚大学的 Chris Upton 开发的（http://athena.bioc.uvic.ca/pgr/vgol/）。

目前其他一些专门用于病毒研究的数据库已经建立（见表 13-3）。EMBL 病毒结构资源数据库提供病毒结构图像，Virgo 网站提供了一套软件工具，其中包括 Viral Genome Organizer 用于图形化显示病毒基因组序列的软件（Upton 等，2000）（图 13-19）。这个网站还包括一个病毒基因组数据库（VGDB），储存有病毒基因组的分析数据（如 GC 含量）等。斯坦福大学 HIV 研究中心和蛋白酶序列数据库提供了一个检索算法，当一个病毒 DNA 序列被提交后，检索结果会指出病毒基因可能发生哪些突变以及这些突变是否会对药物产生抗性（Rhee 等，2003）（图 13-20）。

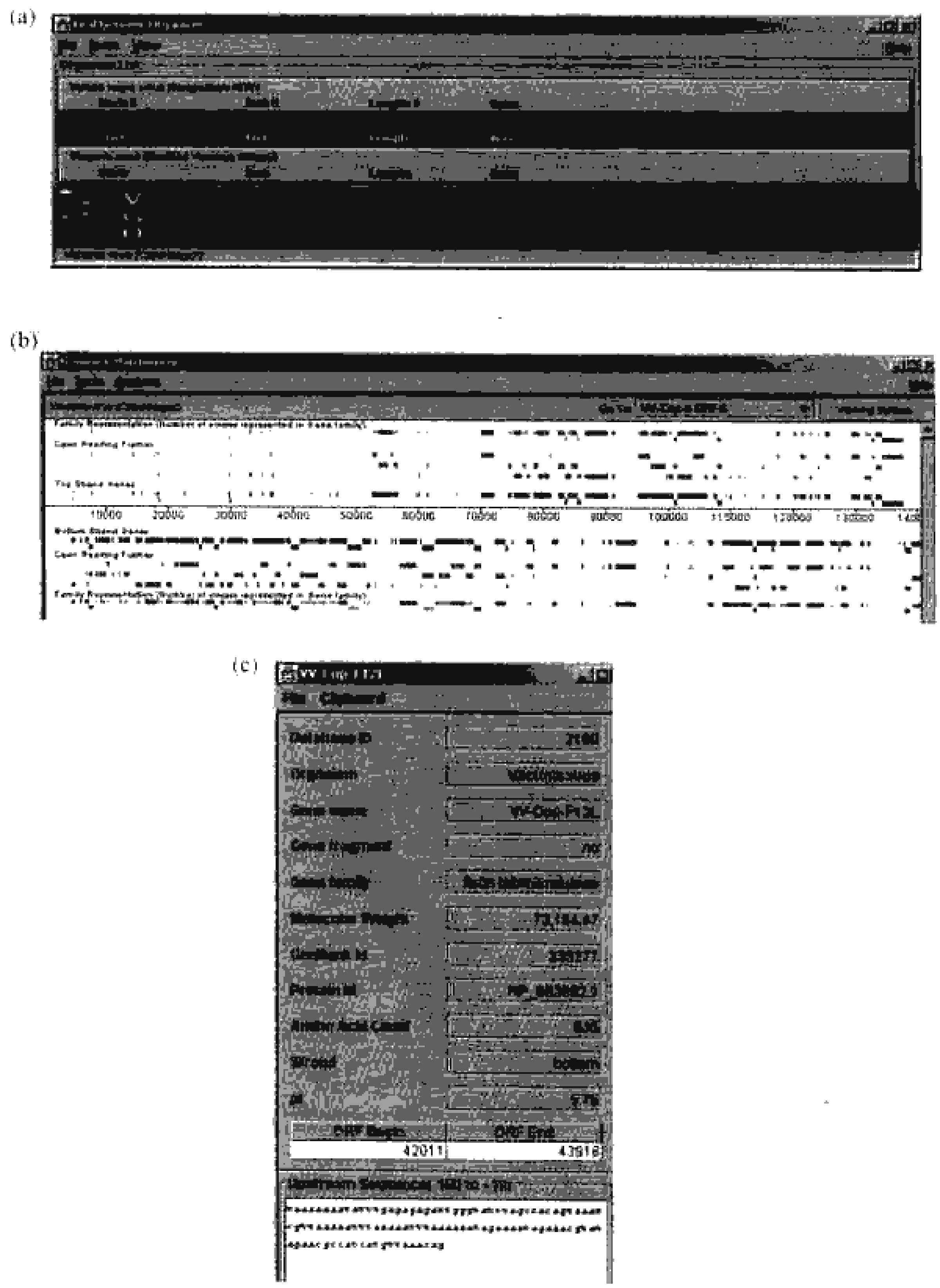


图 13-19 (a) Viral Genome Organizer 是一个基于 Java 的用于分析基因组的工具软件，可用于 30 多种病毒（主要是痘病毒）基因组序列的分析。(b) 经过软件注释的序列。(c) 每一个基因都可通过点击获得更多细节信息

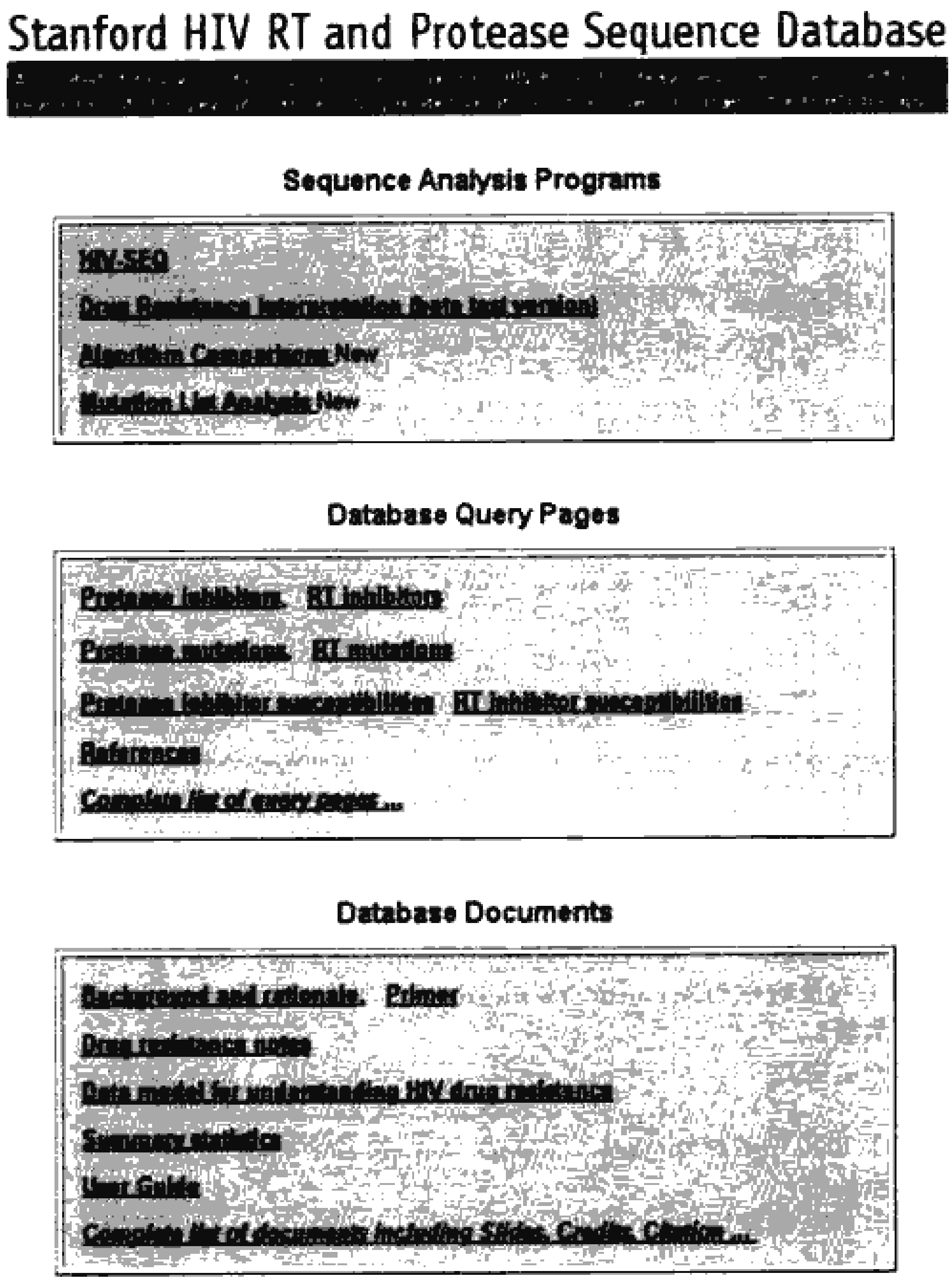


图 13-20 斯坦福 HIV 研究中心数据库（网址：http://hivdb.stanford.edu/hiv/）主要用于对病毒逆转录酶和蛋白酶序列的深入研究

13.4 展望

目前已知的病毒只有几千种，而已知的细菌、古细菌却多达几千万至几亿种（第 14 章），已知的真核物种也有几千万种（第 15～第 16 章）。这种数量上的差别可能是因为病毒需要在宿主细胞中复制繁衍的特殊需要所致。

从本质上来说，那些应用于真核和原核蛋白及核酸序列的生物信息学方法同样也适用于病毒（Kellam，2001）。

BLAST、PSI-BLAST 以及其他数据库搜索工具可能用于寻找病毒序列与其他生物大分子的同源性。

微阵列技术已经用于病毒基因的研究，它能够用于在病毒不同生命周期阶段测量病毒基因的转录水平。

多种相互独立的、用于测量宿主细胞基因转录水平对病毒侵染响应的方法正在发展。

将结构基因组学方法应用于病毒将使得我们能够识别病毒蛋白的三维结构。一些病毒蛋白在药物抑制剂存在下的结构已经被解出。在 NCBI 数据库的 Entrez 蛋白质结构部分已经收录有超过 1700 个病毒蛋白结构的记录。

13.5 常见问题

病毒进化速度很快，一个主要的原因是病毒的核酸聚合酶在复制或转录过程中出错率较高。正因为如此，一个感染了 HIV 的患者，其体内可能携带有几百万个不同的病毒，每一个病毒的 RNA 序列都与其他个体不同。因此，为一些病毒给出“规范”的序列是非常困难的。这使得试图对病毒进化及其基因功能的研究趋向复杂化。

通过对 rRNA 或其他生物分子的序列进行比较得出的生命的系统发生树已经比较完备，然而在这个发生树中却几乎没有病毒的位置。这是因为没有某一种蛋白或基因是所有病毒以及其他生物都拥有的。

13.6 网络资源

网上病毒研究的相关资源见表 13-3。

表 13-3 网上病毒研究的相关资源

资 源	描 述	网 址
ICTVdB	世界病毒数据库	http://www.ncbi.nlm.nih.gov/ICTVdb/Ictv/fr-index.htm
All the Virology on the WWW	提供了大量病毒相关链接和资源	http://www.tulane.edu/~dmsander/garryfavweb.html
The Big Picture Book of Viruses	一般病毒资源	http://www.virology.net/Big-Virology/BVHomePage.html
VIrus Particle ExploreER (VIPER)	提供了 PDB 中的高分辨率病毒结构	http://mmtsh.scripps.edu/viper
Project VirgO	用于分析病毒基因组的软件工具	http://athena.bioc.uvic.ca/genomes/index.html
Viral Genome Organizer	分析大痘病毒和其他病毒	http://athena.bioc.uvic.ca/pbr/vgo
The EMBL Virus Structure Resource	提供了病毒结构图像	http://www.embl-heidelberg.de/ExternalInfo/fuller/icos0.html
Institute for Molecular Virology	位于威斯康星-麦迪逊大学的一个研究所	http://virology.wisc.edu/IMV/about.html
Taxonomy of Plant Viruses	病毒分类	http://www.scri.sari.ac.uk/TiPP/ICTV/ictvhome.html
Stanford HIV RT and Protease Sequence Database	提供药物靶标信息的医学数据库	http://hivdb.stanford.edu/hiv/
INDEX VIRUM	病毒分类	http://life.anu.edu.au/viruses/Ictv/fr-index.htm

13.7 问题讨论

1. 为什么不存在一个包括全部病毒的全面的系统发生树？

2. 如果想构建一个尽可能全面完整的系统发生树，可供选择的有 GenBank 中的 DNA、RNA 和蛋白质序列，你会选择哪一种分子的序列作为参考标准？

13.8 问题

1. 在 LocusLink 数据库中有多少 HIV-1 病毒的蛋白？考虑到 HIV-1 具有极大的各异

性，你也许会预计它的每一个蛋白都有多种变体，那么实际上有多少蛋白序列获得了 RefSeq 编号呢？

2. 在 LocusLink 数据库中找到一条由 RefSeq 编号的 HIV-1 蛋白（例如 Vif 蛋白，NP _ 057851；这只是一个示例，你应该另外找一条）。提交这个蛋白序列做一次 blastp 搜索，使用物种分类报告仔细检查搜索结果。然后重复搜索一次，但是这一次滤掉有关 HIV 的搜索结果。接着上面的示例，你可以在高级搜索选项中输入“vif NOT txid11676 [Organism]”或“vif NOT Hiv [Organism]”来达到这个目的（注意：你可以使用 NCBI 物种分类浏览器来获得 HIV 的物种分类号 txid11676 [Organism]）。从搜索结果看，你所选择的基因或蛋白在各种病毒中是否有广泛的分布？你会觉得可能有某些基因是 HIV 独有的，而其他一些基因在各种病毒中都广泛存在吗？

13.9 自测题

1. 目前已知病毒约几千种，而已知的原核生物和真核生物都在几千万种以上。对于这种数量差异，最令人信服的解释是

(a) 还有大量病毒尚未发现，我们还没有掌握探测它们的方法；

(b) 对大多数病毒而言，我们还没有掌握对它们测序的方法；

(c) 病毒的种类较少是因为它们需要高度独特的生存条件；

(d) 病毒的遗传密码与其他生物不同

2. HIV 病毒基因组包含 9 个蛋白编码基因。GenBank 中这 9 个基因的编号数量大约为

(a) 9；

(b) 900；

(c) 9000；

(d) 90000

3. 运用功能基因组手段分析病毒，我们有可能测量到下列哪种基因表达数据

(a) 病毒侵染人类组织时的病毒基因表达；

(b) 病毒侵染人类组织时的宿主细胞基因表达；

(c) 病毒侵染人类组织时，同时测量宿主细胞和病毒的基因表达；

(d) 病毒侵染人类组织时，分别测量宿主细胞和病毒的基因表达

4. 麻疹病毒最早出现于大约

(a) 2 亿年前；

(b) 200 万年前；

(c) 2 万年前；

(d) 200 年前

5. HIV 病毒最早出现于大约

(a) 7000 万年前；

(b) 700 万年前；

(c) 7000 年前；

(d) 70 年前

6. HIV 病毒亚类型的系统演化

(a) 证实 HIV 病毒是由牛病毒衍生而来；

(b) 可以用于指导开发针对由病毒祖先遗传下来的蛋白序列的疫苗；

(c) 证实哪些人类组织最容易遭受病毒侵染

7. 专门的病毒数据库（如 Oak Ridge 国家实验室数据库）提供了许多在 NCBI 或 EBI 上访问不到的资源，例如

(a) 每一个 HIV 基因的数千个变体的列表；

(b) 前一周发表的关于病毒研究的文献和引用；

(c) 病毒基因组的图形化显示；

(d) 对 HIV 变体在世界各地分布的描述

13.10 推荐读物

Kellam（2001）对生物信息学和功能基因组学方法在病毒学领域的应用的综述。在综述中他描述了从病毒基因结构到蛋白质结构，从病毒基因表达到宿主基因表达的广阔的研究课题。

Flint 等（2000）撰写的非常权威的病毒学教科书。Schaechter 等（1999）撰写的关于微生物疾病机理的专著。这两本书籍中从普通的和临床的视角介绍了种类繁多的病毒。

参 考 文 献

Ackermann, H. W. Frequency of morphological phage descriptions in the year 2000. *Arch. Virol.* **146,** 843–857 (2001).

Barrett, T., and Rossiter, P. B. Rinderpest: The disease and its impact on humans and animals. *Adv. Virus Res.* **53,** 89–110 (1999).

Büchen-Osmond, C. Further progress in ICTVdB, a universal virus database. *Arch. Virol.* **142,** 1734–1739 (1997).

Chang, Y., et al. Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* **266,** 1865–1869 (1994).

Cieslak, T. J., Christopher, G. W., and Ottolini, M. G. Biological warfare and the skin II: Viruses. *Clin. Dermatol.* **20,** 355–364 (2002).

Davison, A. J. Evolution of the herpesviruses. *Vet. Microbiol.* **86,** 69–88 (2002).

Flint, S. J., Enquist, L. W., Krug, R. M., Racaniello, V. R., and Skalka, A. M. *Principles of Virology Molecular Biology, Pathogenesis, and Control.* American Society for Microbiology Press, Washington, DC, 2000.

Flores, R. A naked plant-specific RNA ten-fold smaller than the smallest known viral RNA: The viroid. *C. R. Acad. Sci. III* **324,** 943–952 (2001).

Gaschen, B., Kuiken, C., Korber, B., and Foley, B. Retrieval and on-the-fly alignment of sequence fragments from the HIV database. *Bioinformatics* **17,** 415–418 (2001).

Gaschen, B., et al. Diversity considerations in HIV-1 vaccine selection. *Science* **296,** 2354–2360 (2002).

Hahn, B. H., Shaw, G. M., De Cock, K. M., and Sharp, P. M. AIDS as a zoonosis: Scientific and public health implications. *Science* **287,** 607–614 (2000).

Hendrix, R. W. Evolution: The long evolutionary reach of viruses. *Curr. Biol.* **9,** R914–917 (1999).

Johnson, R. T., and Gibbs, C. J., Jr. Creutzfeldt–Jakob disease and related transmissible spongiform encephalopathies. *N. Engl. J. Med.* **339,** 1994–2004 (1998).

Johnson, C. E., et al. Measles vaccine immunogenicity and antibody persistence in 12 vs 15-month old infants. *Vaccine* **18,** 2411–2415 (2000).

Kellam, P. Post-genomic virology: The impact of bioinformatics, microarrays and proteomics on investigating host and pathogen interactions. *Rev. Med. Virol.* **11,** 313–329 (2001).

La Scola, B., et al. A giant virus in Amoebae. *Science* 299, 2033 (2003).

Mayo, M. A., and Pringle, C. R. Virus taxonomy—1997. *J. Gen. Virol.* **79,** 649–657 (1998).

McClure, M. A. The complexities of genome analysis, the Retroid agent perspective. *Bioinformatics* **16,** 79–95 (2000).

McGeoch, D. J., Cook, S., Dolan, A., Jamieson, F. E., and Telford, E. A. Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. *J. Mol. Biol.* **247,** 443–458 (1995).

Meissner, C., and Coffin, J. M. The human retroviruses: AIDS and other diseases. In M. Schaechter, N. C. Engleberg, B. I. Eisenstein, and G. Medoff (Eds.), *Mechanisms of Microbial Disease.* Lippincott Williams & Wilkins, Baltimore, MD, 1999, Chapter 38.

Paulose-Murphy, M., et al. Transcription program of human herpesvirus 8 (kaposi's sarcoma-associated herpesvirus). *J. Virol.* **75,** 4843–4853 (2001).

Poole, L. J., et al. Altered patterns of cellular gene expression in dermal microvascular endothelial cells infected with Kaposi's sarcoma-associated herpesvirus. *J. Virol.* **76,** 3395–3420 (2002).

Prusiner, S. B. Prions. *Proc. Natl. Acad. Sci. USA* **95,** 13363–13383 (1998).

Reid, A. H., Fanning, T. G., Janczewski, T. A., McCall, S., and Taubenberger, J. K. Characterization of the 1918 "Spanish" influenza virus matrix gene segment. *J. Virol.* **76,** 10717–10723 (2002).

Rhee, S. Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., and Shafer, R. W. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* **31,** 298–303 (2003).

Russo, J. J., et al. Nucleotide sequence of the Kaposi sarcoma-associated herpesvirus (HHV8). *Proc. Natl. Acad. Sci. USA* **93,** 14862–14867 (1996).

Schaechter, M., Engleberg, N. C., Eisenstein, B. I., and Medoff, G. *Mechanisms of Microbial Disease.* Lippincott Williams & Wilkins, Baltimore, MD, 1999.

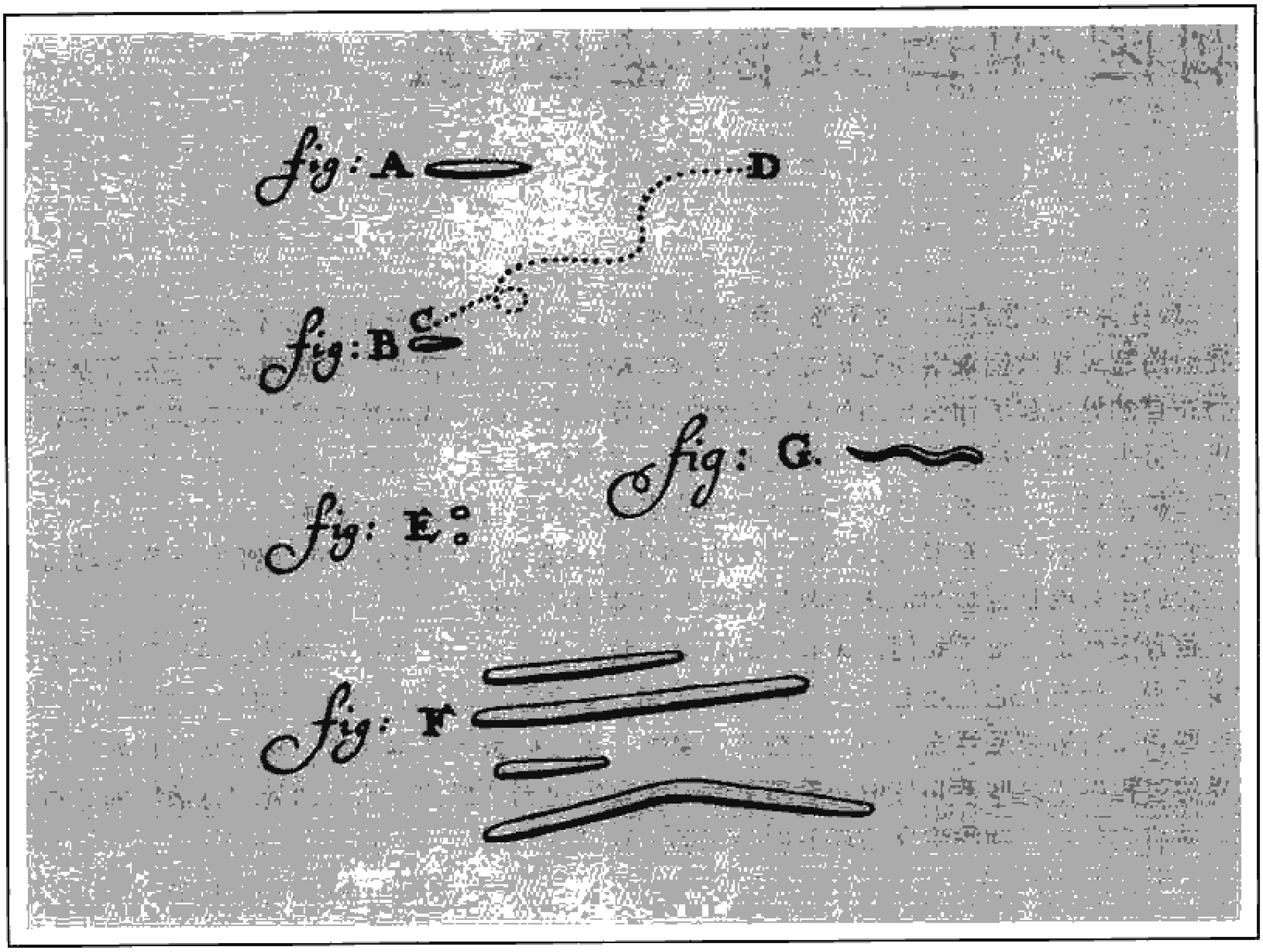
Sharp, P. M., et al. The origins of acquired immune deficiency syndrome viruses: Where and when? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **356,** 867–876 (2001).

Simpson, G. G. The meaning of taxonomic statements. In S. L. Washburn (ed.). *Classification and Human Evolution.* Aldine Publishing Co., Chicago, 1963, pp. 1–31.

Suarez, D. L., et al. Comparisons of highly virulent H5N1 influenza A viruses isolated from humans and chickens from Hong Kong. *J. Virol.* **72,** 6678–6688 (1998).

Upton, C., Hogg, D., Perrin, D., Boone, M., and Harris, N. L. Viral genome organizer: A system for analyzing complete viral genomes. *Virus. Res.* **70,** 55–64 (2000).

Wakeling, M. N., Roy, D. J., Nash, A. A., and Stewart, J. P. Characterization of the murine gammaherpesvirus 68 ORF74 product: A novel oncogenic G protein-coupled receptor. *J. Gen. Virol.* **82,** 1187–1197 (2001).



Antony van Leeuwenhoek（1622～1723）被称作原生动物学与细菌学之父。这张图中所展现的是从他自己口腔中取出的细菌。图A是一种具备运动能力的杆菌；图B是生痰新月形单胞菌，图C和图D显示了它的运动轨迹；图E所示的是两个微球菌；图F表示的是口腔纤毛菌；图G表示的是一个螺旋菌。他在1683年9月17日的一封信中这样描述他所观察到的这些“运动体”：“当我与一个老人交谈时（这个老人过着稳重而节制的生活，从不喝白兰地或抽烟，即便是葡萄酒也极少摄入），我注意到他的牙齿已全部被牙石或牙斑覆盖。于是我问他最近一次清洁口腔是什么时候，他回答说他一生未曾清洁过口腔。我从他口腔中取了一点唾液作检查，然而其中确没有任何曾在我自己或他人唾液中发现的那些东西。所以我又从他的牙缝中取出了一些物质，并分别与他的唾液和洁净的水（其中不含任何“运动体”）混合，这次我观察到了多得不可思议的活“运动体”，比我过去看到过任何人的都多得多。其中数量最多的一类曲着身体前进，如图G所示。此外，其他类型的“运动体”也数量巨大，以至于所有的液体看起来都像是活的（尽管我从他齿间只取了极少的物质与水混合）。”（经Dobell从荷兰语翻译，1932，第242～第243页），授权使用

第 14 章

已完成测序的基因组：细菌和古细菌基因组

“现在你会乐意去问：对于那些用肉眼观察太小的生物的分析讨论什么时候才是尽头，在我们用超级显微镜观察到它们之前，应该忽略哪些东西的存在？这些明显不是很重要的微生物在自然界中起了什么作用使得我们对它们发生了兴趣，或者回报我们在掌握它们常识中付出的努力？我会尽量简单地回答这些问题。多胃纤毛虫，尽管它们相当微小，仍然分担了很多维护自然生态的重要职责，虽然这些或多或少对我们的健康有影响。

它们无处不在，有着难以置信的数目和无法满足的贪食特性，并且腐蚀蔬菜和动物尸体，它们生来就是吞噬和消化吸收这些蔬菜和动物尸体的。

很肯定地说，对于我们健康有利的环境，我们必须在一定程度上感激这些几乎看不见的食腐生物。而且，它们还起了一个更大的作用，就是预防目前地球上有机物质总量的减少。在这些物质溶解或悬浮在水中，被粉碎、腐蚀并最终降解为基本元素，从而从有机界返回无机界的过程中，这些不知疲倦的看不见的“警察”时刻“抓捕”那些四处“逃散”的有机颗粒，并将它们送回动物生命循环的向上流动过程。”

——Richard Owen（1843，第 27 页）

14.1 引言

在本章中我们将会把生物信息学方法运用到生物的三个主要分支中的两个中去：细菌和古细菌。细菌和古细菌被划分为同一组是因为它们都属于原核生物，也就是没有细胞核的单细胞生物。细菌和古细菌也被称为微生物。致病微生物是指那些对人类产生疾病的微生物，包括很多像真菌和原生动物（第 15 章和第 16 章）的真核生物和一些原核生物。

据估计细菌占地球生物总量的 60%。细菌遍布了这个星球上每一个可以想像的角落，大概有 200 万～300 万不同种类的细菌（Fraser 等，2000）。绝大多数细菌和古细菌（超过 99%）都没有被发现和研究过。研究细菌的一个强迫性的缘由就是它们中很多都能使人类和其他动物致病。

在第12章中，我们根据时间顺序叙述了许多细菌和古细菌的基因组测序项目，是从1995年的 *Haemophilus influenzae*（流感嗜血杆菌）开始的。我们在这里将会根据6种不同的标准来给细菌和古细菌进行分类：①形态学；②基因组大小；③生活方式；④和人类疾病的相关性；⑤用 rRNA 进行分子系统发生分析；⑥基于其他分子的分子系统发生分析。我们将会介绍一些研究细菌和古细菌基因组的重要资源：NCBI 和 TIGR 的综合微生物资源库（CMR）。为了更好地描述这些资源，我们将会采用一些生物信息学工具来分析单个的微生物基因组和两个或多个基因组的比较工具。通过比较基因组学，我们开始发现一些微生物学的重要原理，如微生物对特殊生态区域的适应、微生物之间基因的水平转移、基因组扩张和减缩以及致病原因的分子基础。

14.2 根据形态学标准的细菌分类

大部分细菌被分到下面四个大类中：革兰阳性球菌、革兰阴性球菌、革兰阳性杆菌和革兰阴性杆菌（Schaechter，1999）。表14-1是不同种类细菌的例子。革兰着色剂能被大约一半种类的细菌吸收，然后通过它表明细胞壁中蛋白质和肽聚糖的位置。一些其他的细菌有着比较特殊的形状或者染色方式，因此不能按照革兰阳性或革兰阴性球菌或杆菌这种分类方法分类，如莱姆病的致病原伯氏疏螺旋菌（*Borrelia burgdorferi*），此类螺旋菌有着特殊的膜鞘、原生细胞柱和质膜鞭毛（Charon 和 Goldstein，2002）。

表 14-1 基于形态学标准的细菌分类

类　型	例　子①
革兰阳性球菌	*Streptococcus pyogenes*, *Staphylococcus aureus*
革兰阳性杆菌	*Corynebacterium diphtheriae*, *Bacillus anthracis*（炭疽热），*Clostriduium botulinum*
革兰阴性球菌	*Neisseria*, *Gonococcus*
革兰阴性杆菌	*Escherichia coli*, *Vibrio cholerae*, *Helicobacter pylori*
其他	*Mycobacterium leprae*（麻风病），*Borrelia burgdorferi*（莱姆疾病），*Chlamydia trachomatis*（性传播疾病），*Mycoplasma pneumoniae*

① 产生的疾病表示在括号中。

基于分子系统发生的微生物分类如下面描述的一样非常容易理解。分子差别能够揭示种间（表明了生物原核进化树的宽度）和种内（表明了致病菌和与它关系很近但无毒的细菌之间的分子差异）的微生物差异性程度。然而，除了分子标准，还有很多其他基于显微镜和生理研究的方法来区分细菌。例如，区分那些能够进行有氧光合作用（*Cyanobacteria*，藻青菌）或者能够产生甲烷的微生物。

14.3 基于基因组大小和排列的细菌和古细菌分类

在如细菌和古细菌这类的单细胞生物中，基因组大小（或称 *C* 值）就是基因组中 DNA 的总量。细菌和古细菌基因组大小可以在25倍［从0.5～10Mb多（表14-2）］范围内变化（Casjens，1998）。图14-1列出了23个已命名的主要细菌属和它们的亚属的基因组大小。从图中看出，尽管有一些细菌基因组是线性的，但是大多数是环状的；一些细菌基因组由多个环状染色体组成。在大多数细菌属中能发现质粒（小的环状非染色体元件），但是线性的非染色体元件非常少。

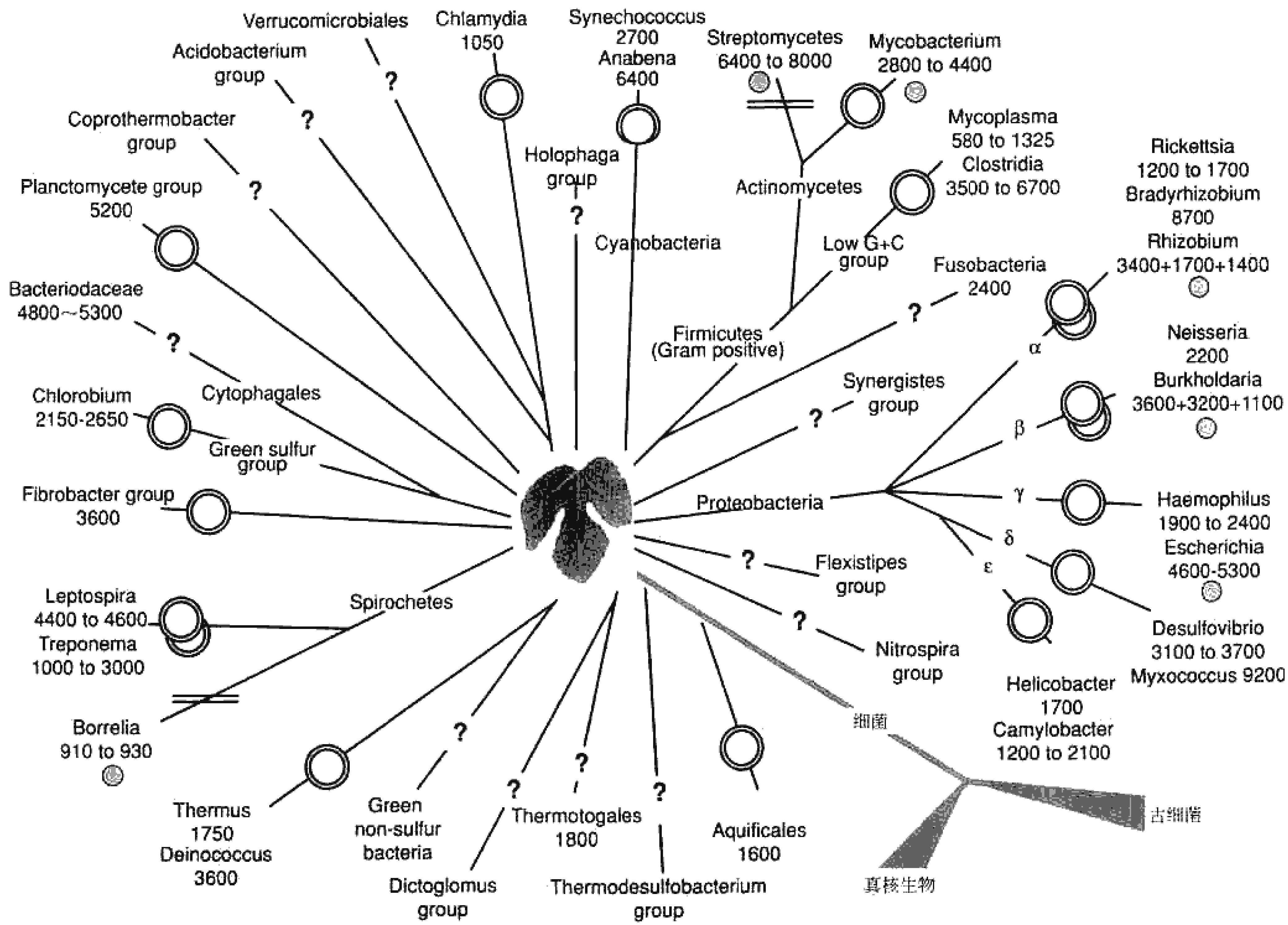


图 14-1 细菌染色体大小和形状。图中显示了23个已经被命名的主要细菌属和它们的一些亚组。这个发生树是基于rRNA序列的，并且是无根的。分支的长度并不代表系统发生距离，中间的叶子表示不确定的分支模式。每个分支的末尾是染色体形状（环形或者线形，有些情况下是多染色体）。还给出了有代表性的属的染色体大小（千个碱基）。在螺旋体菌和放射菌类中普遍存在的线性染色体外元件也显示出来了。这张图是根据Casjens（1998）的修改得到的。授权使用

表 14-2　细菌和古细菌基因组大小范围

分　类	基因组大小/Mb	比例(最高/最低)
细菌	0.58～13.2	23
Mollicutes	0.58～2.2	4
革兰阴性菌	0.65～9.5	15
革兰阳性菌	1.6～11.6	7
蓝青菌	3.1～13.2	4
古细菌	0.5～4.1	3

注：改编于 Graur 和 Li（2000，第 36 页）授权使用。

裂殖酵母 *S. pombe* 的基因组大小为 13Mb，编码了大约 4800 个基因（Wood 等，2002）（第 15 章）。一些细菌基因组比真核生物基因组大得多。天蓝色链霉菌（*Streptomyces coelicolor*）基因组是目前为止已完成测序的最大细菌基因组（8.7Mb），编码 7800 个蛋白质（Bentley 等，2002）。

在双倍体或四倍体生物中，基因组大小指没有复制的单倍体基因组的 DNA 大小。我们将会在第 15～第 17 章讨论真核生物基因组的大小。

总的来说，细菌基因组的编码基因数目大概在 500～8000。对于许多已经完成基因组测序的细菌，蛋白质编码基因占了整个基因组的 85%～94%。这样，基因间部分和非基因部分就很少了［有一个例外是麻风病的病原体——麻风杆菌（*Mycobacterium leprae*），它的基因组发生了巨大的基因丢失，蛋白质编码基因只占整个基因组的 49.5%（Cole 等，2001）］。微生物基因组的基因密度基本上都是每 1000 个碱基一个基因。即使在生殖支原体（*Mycoplasma genitalium*）基因组这样非常小的基因组中，基因组大小与基因密度和基因平均大小没有关系（Fraser 等，1995）。表 14-3 列出了一些较大或较小的细菌和古细菌基因组大小。

非寄生生物中最小的已测序基因组是生殖支原体（*Mycoplasma genitalium*）——一种泌尿生殖器病原体。*M. genitalium* 有 580070 个碱基对，编码了 470 个蛋白质编码基因、3 个 rRNA 基因和 33 个 tRNA 基因（Fraser 等，1995）。支原菌是柔膜体纲类的细菌，它们没有细胞壁，并且 GC 含量比较低（32%）。

另外一个例外是寄生虫 *Rickettsia prowazekii*，它含有 24% 的非编码 DNA。

表 14-3　基因组相对较大或者较小的细菌和古细菌基因组大小

物　种	基因组大小/Mb	编码区	GC 含量	参考文献
Streptomyces coelicolor	8.67	7825	72	Bentley 等，2002
Burkholderia cepacia	7.6(3.7+3.1+0.8)		65	未出版
Pseudomonas fluorescens	6.6	NA	66	未出版
Mycobacterium marinum	6.0	NA	65	未出版
Tropheryma whippelii	0.93	808	46	未出版
Ureaplasma urealyticum parvum biovar serovar 3	0.752	613	26	Glass 等，2000
Mycoplasma pneumoniae M129	0.816	677	40	Himmelreich 等，1996
Mycoplasma genitalium G-37	0.58	470	32	Fraser 等，1995
Nanarchaeum equitans	0.5	NA	NA	Huber 等，2002

注：1. 改编自 http://www.sanger.ac.uk/Projects/Microbes/ 和 NCBI 网站（PubMed，Entrez）。
2. NA 表示未获得。

我们介绍了一些在 NCBI 网站上查看一个细菌基因组（如 *M. genitalium*）信息的方法（第 12 章）。另外一个重要的网站是 TIGR 的 CMR。CMR 对已经完成的微生物基因组进行

了注释（Peterson 等，2001）。*M. genitalium* 的记录如图 14-2 所示，包括到 PubMed 的链接、该生物的背景信息、下载核苷酸或蛋白质序列的 ftp 站点和染色体图。染色体图中包含一个具有基因组注释的可点击图谱（图 14-3）。CMR 提供了许多用来分析原核基因组的工具，我们将在本章下文讨论。

为了访问 CMR，登录 TIGR（http://www.tigr.org），然后点击 CMR 链接。

现在已经在极端嗜热古菌中发现了一个基因组比 *M. genitalium* 基因组更小的生物—— *Nanoarchaeum equitans*，它生长在海底火上口附近（Huber 等，2002）。这种生物附着在另外一种古细菌（*Ignicoccus*）上。由于它的细胞和基因组极小，Huber等（2002）称 *N. equitans* 可能是一种介于最小的活性生物（如 *M. genitalium*）和大病毒（如天花病毒）之间的中间体。然而，寄生于细胞内的细菌和古细菌与病毒不同，它们被划分为非共生生物。通过与小原核基因组比较，可能可以用来估计生物所需要的最小基因数目（框 14-1）。

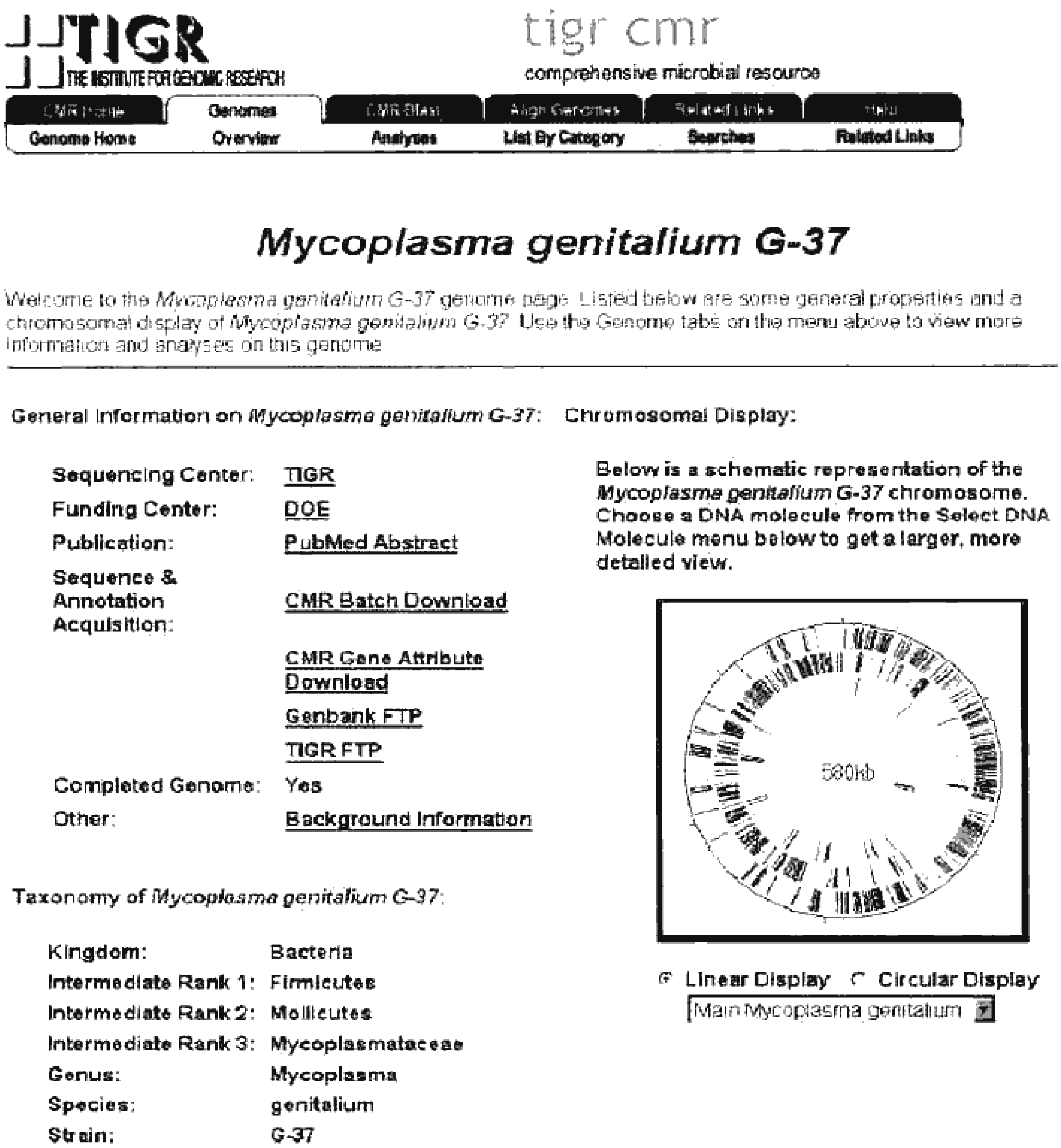


图 14-2 TIRG 的 CMR 为微生物的基因组研究提供了一个最重要的网站（http://www.tigr.org）。这是一个典型的物种显示页面，包括分类描述、文献和背景信息的链接以及注释的基因组序列

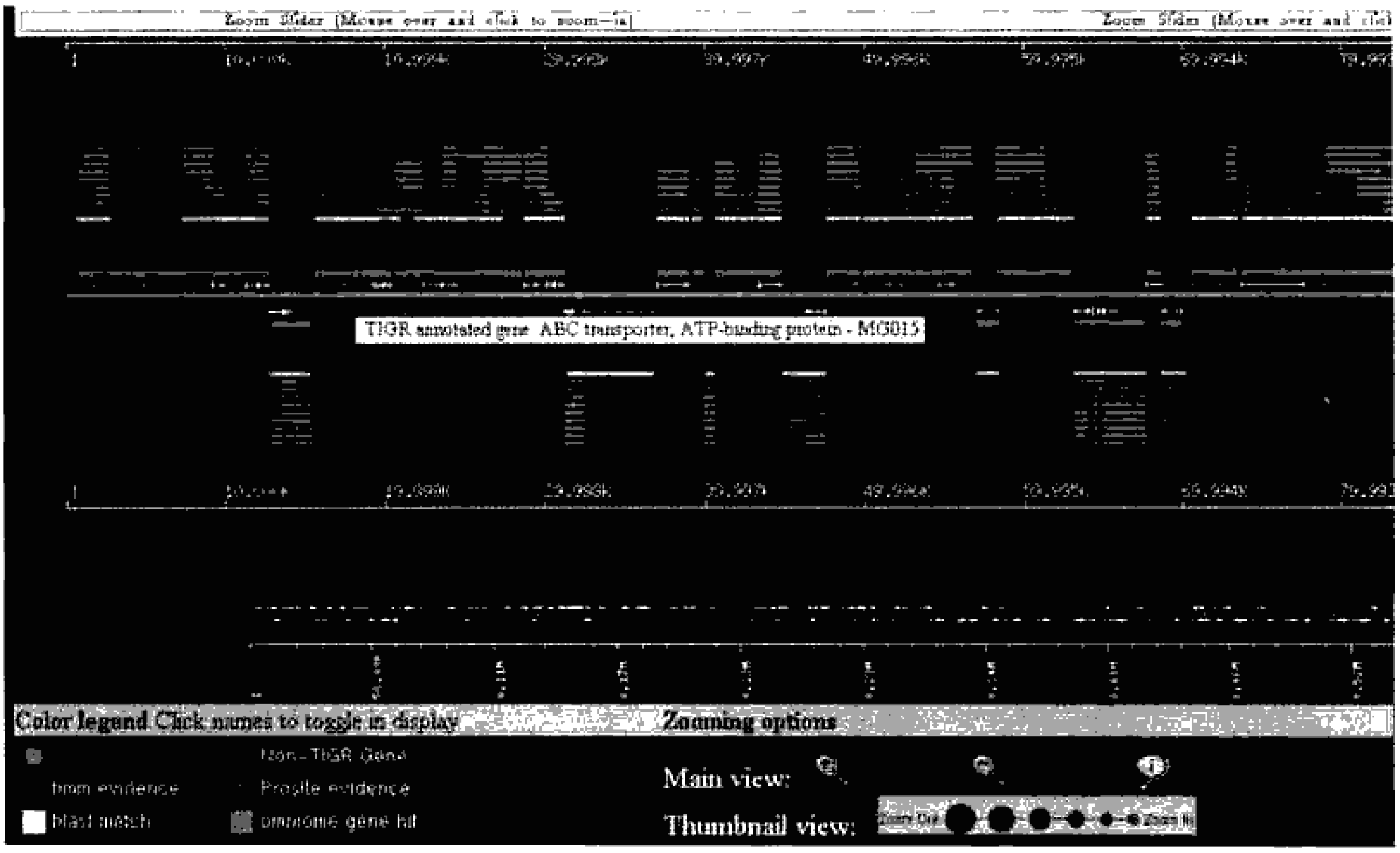


图 14-3　在 TIRG 的 CMR 的输入界面中点击基因组图像，你就能看到数据库中几十个基因组中的任何一个注释基因组

框 14-1　最小的基因组：生命所需最小基因组为多大？

最小的活的生物（即最小的能够自我复制的生物）的基因组需要多少基因？要回答这个问题，一种方法就是找出一些微生物中共有的同源蛋白。Mushegian 和 Koonin（1996）找出了在 *Escherichia coli*、*H. influenzae* 和 *M. genitalium* 共有的 239 个同源基因，这被认为是估计出来的最小的基因组大小。然而，有些基因在功能上非常重要但是在序列水平上并不是非常近源，可能会发生非同源基因替代。Mushegian 和 Koonin 发现的 239 个基因中，主要包括下面几种功能：翻译、DNA 复制、DNA 修复和重组、转录、厌氧代谢、脂类和辅因子的生物合成、跨膜转运。他们还发现了一些不知功能的蛋白质。

值得注意的是，*H. influenzae*（一种革兰阴性菌）和 *M. genitalium*（一种革兰阳性菌）是相当远源的，它们共同的祖先生活于 20 亿年前（Feng 等，1997）。这样，Mushegian 和 Koonin 得到的最小基因列表不能反映它们共同的祖先的基因（Doolittle，1998）。

Itaya 发明了一种找出生命所需的最少基因数目的实验方法（1995），他在细菌 *Bacillus subtilis* 中随意地敲除蛋白质编码基因，其 79 个基因座中只有 6 个基因座的突变能够抑制细菌生长，是不可或缺的。由此可以推得对于 *B. subtilis* 全基因组，估计有大概 250 个基因是必需的。

14.4　基于生活方式的细菌和古细菌分类

除了形态学和基因组大小与排列这两个标准，第三种用来给细菌（和古细菌）分类的方法是基于它们的生活方式。用这种方法来给微生物分类的一个主要的优点就是能够突出基因组减小这一特点，而这个往往是和三种生物（极端微生物、胞内原核生物和表细胞原核生物）的生活方式结合在一起的。

细菌的分类一定程度上是由 Jim Moulder 的在线评论演变来的（http://www.chlamydiae.com/Evolution-index.htm）。

① 极端微生物是生活在极端环境下的微生物。古细菌在一些

像火山口（极端嗜热古菌）、缺氧栖息地（甲烷产物）等高盐环境和地热区域被发现。

② 胞内细菌入侵真核细胞：一个很有名的例子就是 α-proteobacterium，认为它侵入了真核细胞后就进化成了现在的线粒体。

③ 表细胞细菌（和古细菌）是生活在宿主附近但不在宿主细胞内的寄生虫。

我们可以区分出细菌和古细菌的六种基本生活方式（表 14-4）：

表 14-4 基于生态环境的细菌和古细菌分类

生活方式	细菌	基因组大小/Mb	参考文献
胞外	*Escherichia coli*	4.6	Blattner 等，1997
	Vibrio cholerae	4.0	Heidelberg 等，2000
	Pseudomonas aeruginosa	6.3	Stover 等，2000
	Bacillus subtilis	4.2	Kunst 等，1997
	Clostridium acetobutylicum	4.0	Nolling 等，2001
	Deinococcus radiodurans	3.3	White 等，1999
兼性胞内	*Salmonella enterica*	4.8	Parkhill 等，2001a
	Yersinia pestis	4.7	Parkhill 等，2001b
	Legionella pneumophila	3.9	Bender 等，1990
	Mycobacterium tuberculosis	4.4	Cole 等，1998
	Listeria monocytogenes	2.9	Glaser 等，2001
嗜极性	*Aeropyrum pernix*	1.7	Kawarabayasi 等，1999
	Methanococcus janneschi	1.7	Bult 等，1996
	Archeoglobus fulgidus	2.2	Klenk 等，1997
	Thermotoga maritima	1.9	Nelson 等，1999
	Aquifex aeolius	1.6	Deckert 等，1998
细胞表层	*Neisseria meningitidis*	2.2	Tettelin 等，2000
	Haemophilus influenzae	1.8	Fleischmann 等，1995
	Mycoplasma genitalium	0.6	Fraser 等，1995
	Mycoplasma pneumoniae	0.8	Himmelreich 等，1996
	Ureaplasma urealyticum	0.8	Glass 等，2000
	Mycoplasma pulmonis	1.0	Chambaud 等，2001
	Borrelia burgdorferi	0.9	Fraser 等，1997；Casjens 等，2000
	Treponema pallidum	1.1	Fraser 等，1998
	Helicobacter pylori	1.7	Tomb 等，1997；Alm 等，1999
	Pasteurella multocida	2.3	May 等，2001
胞内专性，共生	*Buchnera* sp.	0.6	Shigenobu 等，2000
	Wolbachia spp.	1.1	Sun 等，2001
	Wigglesworthia glossinidia	0.7	Akman 等，2002
	Sodalis glossinidius	2.0	Akman 等，2001
胞内专性，寄生	*Rickettsia prowazekii*	1.1	Andersson 等，1998
	Rickettsia conorii	1.3	Ogata 等，2001
	Ehrlichia chaffeensis	1.2	—
	Cowdria ruminantium	1.6	de Villiers 等，2000
	Chlamydia trachomatis	1.1	Stephens 等，1998；Read 等，2000
	Chlamydophila pneumoniae	1.3	Kalman 等，1999；Read 等，2000；Shirai 等，2000

注：改编自 http://www.chlamydiae.com。

① 细胞外：例如，大肠杆菌 *E. coli* 一般生活在人类肠道内，但是并不进入细胞。

② 胞内细菌能够进入宿主细胞中，但是这种行为也决定于其所处环境。*Mycobaterium tuberculosis*（结核病的病原体）能够在感染巨噬细胞后潜伏着，直到几十年后才被激活致病。

③ 极端微生物：古细菌最初都在极端环境下被发现。一些古细菌生活在温度高达 113℃、pH0、氯化钠浓度高达 5mol/L 的环境下。*Methanococcus janneschi* 是第一个完成基因组测序的古细菌（Bult 等，1996），它生活在压力超过 2.0×10^4 kPa 的环境下，并且它的最适温度为 85℃ 左右。随后古细菌也在一些极端栖息地被发现，如森林和海水中（DeLong，1998）。

每年，世界上有 190 万人死于肺结核，另外有 19 亿人感染（http://www.cdc.gov/ncidod/eid/vol8no11/02-0468.htm）。*M. tuberculosis* 的基因组是由 Cole 等（1998）测序的。

④ 表细胞原核生物生长在它们的宿主之外，但却和宿主有关。*Mycoplasma pneumoniae* 是一种基因组大小约为 8.16×10^5 bp 的细菌，它们是呼吸道感染的主要致病原。该细菌是表面寄生虫，附着在宿主的呼吸道上皮细胞上。其基因组是由 Himmelreich 等在 1996 年完成测序的，然后由 Peer Bork 和他的同事们重新注释。

⑤ 必须是胞内共生的：Tamas 等（2002）比较了 *Buchnera aphidicola*（Sg）和 *Buchnera aphidicola*（Ap）这两种细菌的全基因组序列，这两个细菌分别是蚜虫 *Schizaphis graminum*（Sg）和 *Acyrthosiphon pisum*（Ap）的细胞内共生生物。它们的基因组大小都是 6.4×10^5 bp 左右，属于最小的已知细菌基因组之列。它们分别有 564 个和 545 个基因，其中大部分基因是相同的（526 个）。值得注意的是，这两个细菌是在 500 亿年前发生分支的，然而它们的基因组结构还是具有相当的保守性。从发生分支后它们的基因组中没有发生基因倒转、基因移位、基因复制和增加（Tamas 等，2002）。这是一个基因组不发生演化的例子。这种必须胞内细菌如此强的基因组保守性的例子还是很罕见的。细胞内共生生物的基因组一般都比较小，这也反映了这些细菌主要依靠它们宿主的营养。

⑥ 必须是胞内寄生的：*Rickettsia prowazekii* 是流行性斑疹伤寒的致病菌。它的基因组比较小，只有 1.1Mb（Andersson 等，1998）。像其他立克次体一样，它是一种 α-proteobacterium，只选择性地感染真核细胞。由于其基因组和线粒体基因组非常相近因而也受到很多关注。一个与它非常近源的物种 *Rickettsia conorii* 是导致人类地中海斑疹热的必需胞内寄生菌。Ogata 等人（2001）对它的基因组进行了测序。同 *Buchnera aphidicola* 亚种一样，两种立克次体寄生菌的基因组也是相当保守的。

为什么一些细菌基因组严重减小呢？这是因为胞内寄生虫经常遭受到有害突变和基因替代使得基因丢失的缘故（Andersson 和 Kurland，1998）。从原始 α-proteobacterium 到现在的线粒体的进化过程中也发生了这个过程，现在的线粒体只保留了相当小的基因组大小。

14.5 基于与人类疾病相关的细菌分类

细菌和真核生物已经相互"交战"几百年了。细菌为了繁殖需要占据人体这个营养丰富的环境。典型的细菌"殖民地"包括皮肤、呼吸道、消化道（口腔、大肠）、尿道和生殖系统（Eisenstein 和 Schaechter，1999）。据估计每个人身上的细菌数目超过自身的细胞数目。大多数情况下，这些细菌对人类是无害的。然而，有些细菌能够导致感染，甚至带来灾难性的后果。

世界上每年有690000个新麻风病例报道，其致病病原体就是麻风分枝杆菌。每年有几百万个关于沙门菌病的报道，这是由肠沙门菌导致的。一个大肠杆菌的致病株（O157：H7）会导致出血性大肠炎，每年在美国感染75000人。像上文所提到的，*M. tuberculosis* 感染了几亿人并导致了几百万人死亡。你能够在疾病控制和预防中心的网站上阅读到许多关于细菌疾病的例子。

最近一些年，由于广泛使用抗生素导致了细菌抗药性的泛滥，因此急需找到细菌的毒性因子，然后找到相应的接种疫苗。对这个问题的一个解决办法就是比较细菌的致病株和非致病株（见下文）。表14-5列出了一些通过接种疫苗已经控制的细菌疾病。由细菌导致的世界范围的疾病负担已经相当严重了。

到了2003年中期，已经有大约150个已经完成的微生物基因组测序计划，另外还有200个正在进行中。一些有代表性的人类致病细菌基因组已经被测序了。

原核生物世界一个现行的主题就是细菌菌落基因经过了重组，导致了基因多样化。Joyce等（2002）针对如 *Helicobacter pylori*（胃溃疡的主要致病因子）、*Streptococcus pneumoniae* 和 *Salmonella enterica* 这样的致病细菌验证了这个情况。真核生物经过有性生殖实现了基因多样化，原核生物通过其他一些机制（如重组和水平基因转移）也实现了基因的多样化。Joyce等（2002）阐述了用DNA基因芯片技术来衡量基因表达改变的重要作用，而基因表达改变同导致人类不同疾病表现型的细菌有着密切的关系。

表14-5 能够通过免疫接种预防的细菌疾病

疾病	物种
炭疽热	*Bacillus anthracis*
腹泻病(霍乱)	*Vibrio cholerae*
白喉	*Corynebacterium diphtheriae*
社区获得性肺炎	*Haemophilus influenzae* type B, *Streptococcus pneumoniae*
莱姆病	*Borrelia burgdorferi*
脑膜炎	*Haemophilus influenzae* type B(HIB), *Streptococcus pneumoniae*, *Neisseria meningitidis*
百日咳	*Bordetella pertussis*
破伤风	*Clostridium tetani*
肺结核	*Mycobacterium tuberculosis*
伤寒	*Salmonella typhi*

注：改编自 http://www.cdc.gov/nip/diseases/disease-chart-hcp.htm 和 http://www.cdc.gov/ncidod/dbmd/diseaseinfo/default.htm。

14.6 基于核糖体RNA序列的细菌和古细菌分类

Brochier和Philippe(2002)已经为极端嗜热细菌（例如 *Aquificales* 和 *Thermotaogales*）是最底部的分支这个观点而争论。相反，他们认为 *Planctomycetale* 才在发生树的底部。

我们已知的对微生物多样性抽样调查的主要途径就是分子系统发生分析。根据不同物种的16S rRNA的多序列比对，我们能画出系统发生树。作为系统发生分析分子，核糖体RNA有着其杰出的特性：它在整个细胞中都有分布，高度保守却又有足够的多样性来解释有意义的差别，并且它很少在物种间转移。图14-1是一个例子，我们在图12-1上也能看到相似的发生树。

Carl Woese和他的同事们（Woese和Fox，1977；Fox等，1980）最早对rRNA研究的主要结论就是细菌和古细菌属于不同的

组。发生树上根部的分支是极端嗜热微生物，这一点与生命的祖先生活在高温下这个假设相一致（Achenbach Richter 等，1987）。

我们认识微生物多样性的一个很大的收获就是意识到绝大部分细菌和古细菌都是不能培育的（Hugenholtz 等，1998）。很容易从自然界中获得一些微生物，然后让它们在不同的生长培养基中生长。对于大多数其他微生物，其生长条件目前并不知晓。也可以直接从自然栖息地中抽取核酸来对未培育的（或者不能培育的）微生物进行取样（Delong 和 Pace，2001）。早期对古细菌的研究将它们描述成为嗜极生物，但最近研究表明它们也存在于一些非极端栖息地，如海洋表层海水和农田中（DeLong，1998）。海洋中所有微生物的 20%可能是古细菌（DeLong 和 pace，2001）。

由于取样偏差，发现最多的四种细菌属为：*Proteobacteria*、*Firmicutes*、*Actinobacteria* 和 *Bacteroidetes*（Hugenholtz，2002）。这几个种属的细菌覆盖了超过 90%的已知细菌（Gupta 和 Griffiths 讨论过，2002）。现在已经知道了有 35 个细菌属和 18 个古细菌属（Hugenholtz，2002）。对没有经过培养的微生物的分析能够开阔我们对细菌和古细菌多样性的眼界。

Reysenbach 和 Shock (2002) 根据 16S rRNA 的序列生成了一个嗜极微生物的系统发生树。他们用了一个专门为 rRNA 研究设计的软件包，叫做 ARB。你能够在 http://www.arb-home.de/得到这个软件。

TIGR 的 CMR（见上）列出了细菌的 12 个主要的分类（表 14-6）和古细菌的两个主要分类：嗜泉古菌界（Crenarchaeota）和广域古菌界（Euryarchaeota）（表 14-7）。这个分类方式和 NCBI 的差不多（图 14-4）。

表 14-6　在 TIGR 的 CMR 上的细菌分类

中间物次序 1	中间物次序 2	属、种、株(例子)	基因组大小/Mb	GenBank 编号
Actinobacteria	Actinobacteridae	*Mycobacterium tuberculosis* CDC1551	4.4	NC_002755
Aquificae	Aquificales	*Aquifex aeolicus* VF5	1.5	NC_000918
Bacteroidetes	Bacteroides	*Porphyromonas gingivalis* W83	2.3	NA
Chlamydiae	Chlamydiales	*Chlamydia trachomatis* serovar D	1.0	NC_000117
Chlorobi	Chlorobia	*Chlorobium tepidum* TLS	2.1	NC_002932
Cyanobacteria	Chroococcales	*Synechocystis* sp. PCC6803	3.5	NC_000911
	Nostocales	*Nostoc* sp. PCC7120	6.4	NC_003272
Deinococcus-Thermus	Deinococci	*Deinococcus radiodurans* R1	2.6	NC_001263
Firmicutes	Bacillales	*Bacillus subtilis* 168	4.2	NC_000964
	Clostridia	*Clostridium perfringens* 13	3.0	NC_003366
	Lactobacillales	*Streptococcus pneumoniae* R6	2.0	NC_003098
	Mollicutes	*Mycoplasma genitalium* G-37	0.580	NC_000908
Fusobacteria	Fusobacteria	*Fusobacterium nucleatum* ATCC25586	2.1	NC_003454
Proteobacteria	Alphaproteobacteria	*Rickettsia prowazekii* Madrid E	1.1	NC_000963
	Betaproteobacteria	*Neisseria meningitidis* MC58	2.2	NC_003112
	Epsilon subdivision	*Helicobacter pylori* J99	1.6	NC_000921
	Gamma subdivision	*Escherichia coli* K12-MG1655	4.6	NC_000913
	Magnetotactic cocci	*Magnetococcus* sp. MC-1	NA	NA
Spirochaetales	Spirochaetaceae	*Borrelia burgdorferi* B31	0.91	NC_001318
Thermotogales	Thermotoga	*Thermotoga maritima* MSB8	1.8	NC_000853

注：1. 细菌采用“界”来描述，接下来是“中间物次序”，“中间物次序”没有列出。

2. 参见 http://www.tigr.org/tigr-scripts/CMR2/CMRGenomes.spl，“NA”表示未获得。

表 14-7 在 TIGR 的 CMR 上的古细菌分类

中间次序 1	中间次序 2	属、种、株(例子)	基因组大小/Mb	GenBank 编号
Crenarchaeota	Thermoprotei	*Aeropyrum pernix* K1	1.6	NC_000854
Euryarchaeota	Archaeoglobi	*Archaeoglobus fulgidus* DSM4304	2.2	NC_000917
	Halobacteria	*Halobacterium* sp. NRC-1	2.0	NC_002607
	Methanobacteria	*Methanobacterium thermoautotrophicum* delta H	1.7	NC_000916
	Methanococci	*Methanococcus jannaschii* DSM2661	1.6	NC_000909
	Methanopyri	*Methanopyrus kandleri* AV19	1.6	NC_003551
	Thermococci	*Pyrococcus abyssi* GE5	1.7	NC_000868
	Thermoplasmata	*Thermoplasma volcanium* GSS1	1.5	NC_002689

注：1. 细菌采用"界"来描述，接下来是"中间物次序"，"中间物次序"没有列出。

2. 参见 http://www.tigr.org/tigr-scripts/CMR2/CMRGenomes.spl。

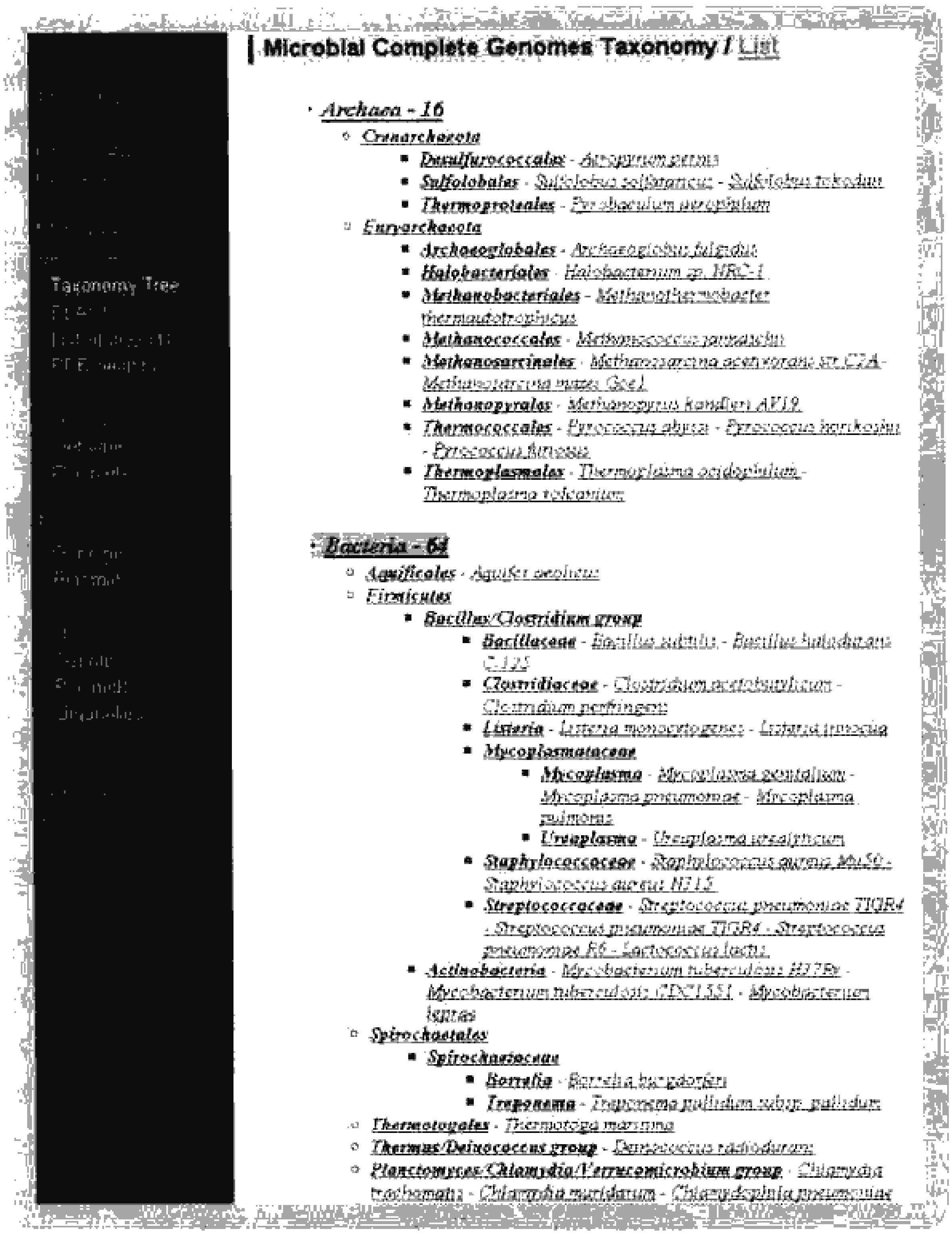


图 14-4 NCBI 上的微生物分类树包括古细菌的两个亚类和细菌的一些亚类（http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/new_micr.html）

14.7 基于其他分子序列的细菌和古细菌分类

除了核糖体 RNA，许多其他 DNA、RNA 或者蛋白质序列都能用来进行分子系统发生

研究。利用单个蛋白质（或基因）产生的系统发生树的拓扑结构经常自相矛盾，并且同用 rRNA 得到的结果也矛盾。这些矛盾通常是由于水平基因转移（见下文，能够混淆系统发生树构建）或者由于基因或蛋白质序列的替代饱和导致的系统发生信号的丢失而导致的。为了避免这个问题，可采用的方法之一就是用联合基因集或联合蛋白质集。Brown 等（2002）比对了在 45 个物种中保守的 23 个同源蛋白。他们得到的系统发生树支持嗜热菌是最早进化的细菌这个观点（图 14-5）。

还有许多其他方法来进行细菌系统发生分析，其中之一就是找出一大组蛋白质中保守的

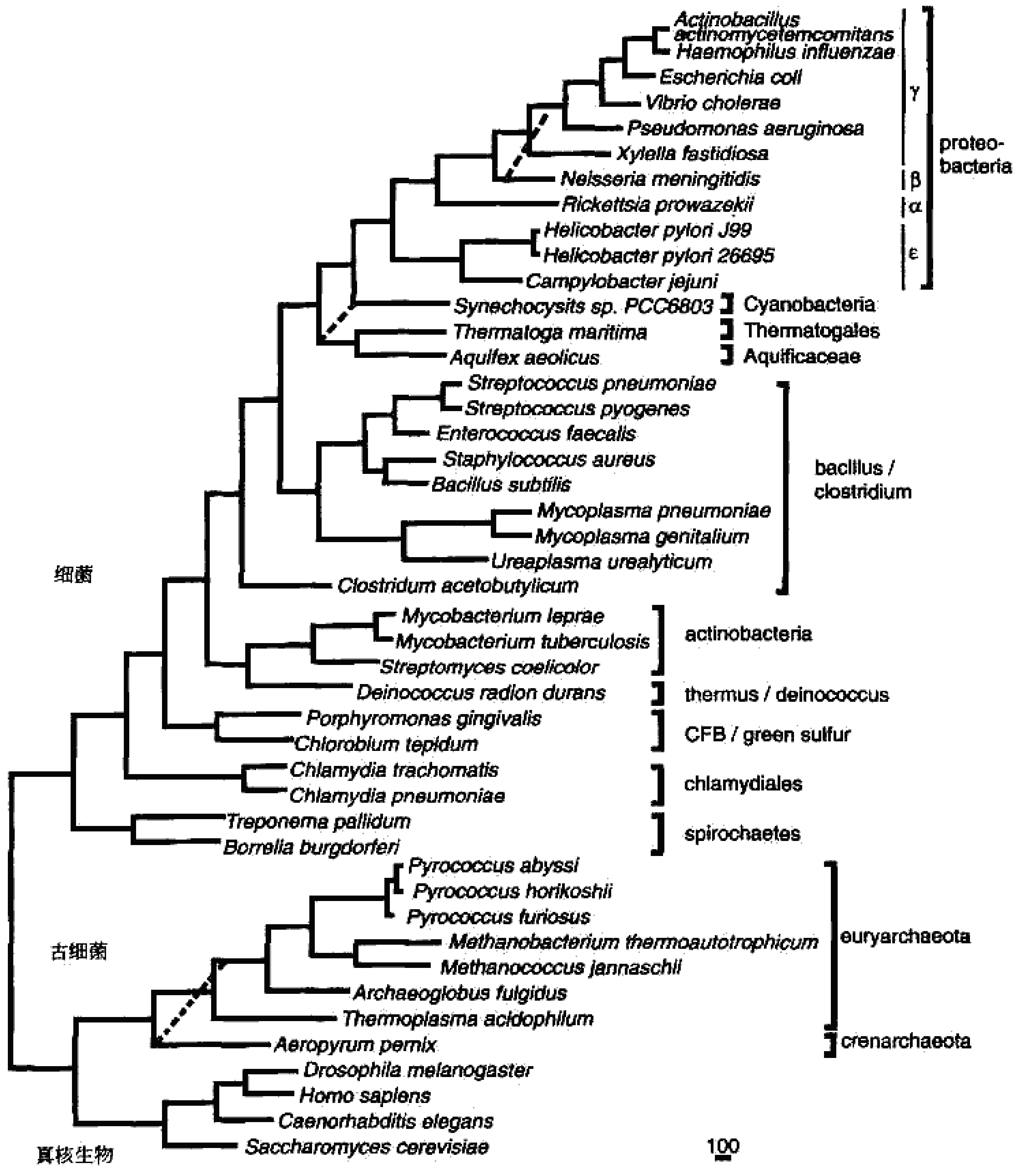


图 14-5 根据 Brown 等（2001）改编的无根树是由 23 个蛋白质（有 6591 个氨基酸残基）比对得到的。这些蛋白质在 45 个物种中都是保守的，包括 tRNA 合成酶、延长因子和 DNA 聚合酶Ⅲ亚单位。有许多关于系统发生信息的网站包含这些蛋白质。这个发生树有 3 个主要的分支，且已经在第 12 章讨论过了。授权使用

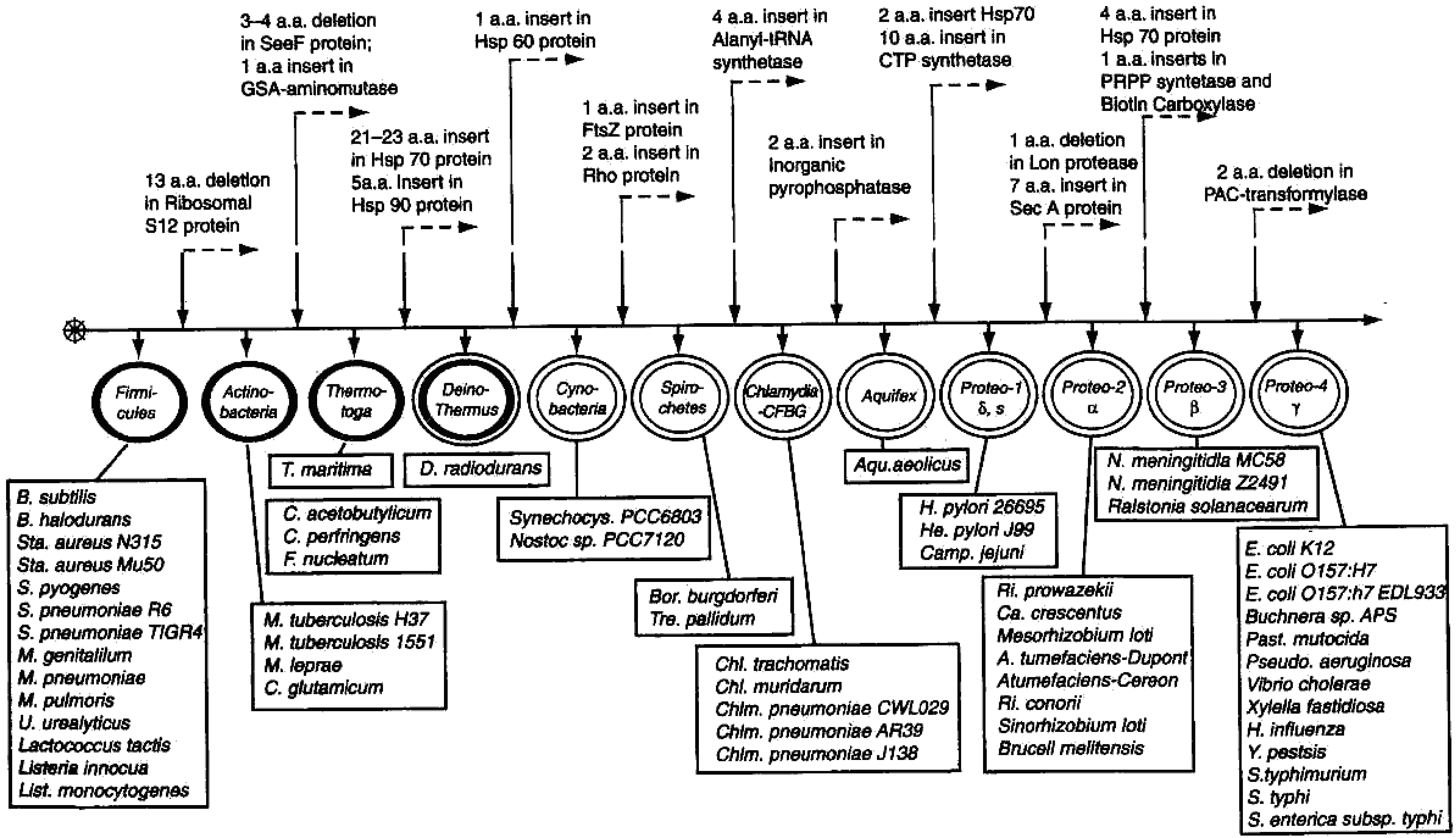


图 14-6 基于不同蛋白质之间的信号序列的细菌种类之间的系统发生关系。线上的实箭头是指可能发生插入或者删除的进化位点。这样，每个箭头的右边所有的细菌组都有这个插入/删除信号序列，而箭头左边的就没有这种变化。来自 Gupta 和 Griffiths（2002），授权使用

插入和删除。这种“签名序列”能够区分出细菌种属并形成发生树的根基（图 14-6）（Gupta 和 Griffiths，2002）。这种发生树从全基因组上显示了细菌种类的相对分支顺序。Eugene Koonin 和他的同事们（Wolf 等，2001）用 5 种独立的方法对 30 个完成测序的细菌基因组和 10 个完成测序的古细菌基因组构建了系统发生树。

同源细菌基因数据库(HOBACGEN) 是一个从 SWISSPROT 和 TrEMBL 来的蛋白质数据库，将蛋白质分成不同的家族来进行系统发生分析(Perriere 等,2000)。你能够在 http://pbil.univlyon1.fr/databases/hobacgen.html 得到这个数据库。

① 他们用 COG 数据库来评估每一个基因组中存在或不存在的基因（见下文）。最后找到了 17 个稳定的基因（所有的都是编码核糖体蛋白质和 RNA 聚合酶亚单位的）。至少有一个基因组中，另外 4586 个基因对丢失了。

② 他们估计了基因组中局部基因顺序的保守性（如相邻基因对）。

③ 他们计算了两个可能的同源基因之间一致率的分布。

④ 他们将 32 个核糖体蛋白质放在一起进行多序列比对（一共有 4821 列），然后用最大似然法产生了一个系统发生树。

我们将会从联合蛋白质数据集的发生树角度来研究真核生物

⑤ 他们比较了一些从一系列蛋白质比对得到的发生树。Wolf 等（2001）得出结论说传统的比对方法和现在基于局部基因顺序的基因组数据分析的新方法有效性差不多。然而，这些新方法能够得到一些不同的信息（例如，对同源蛋白质进行分析能够找到在同一种系中丢失或者发生水平转移的基因）。

14.8 原核基因组分析

14.8.1 核苷酸组成

对已经完成的基因组进行分析，发现核苷酸组成具有一些特征性质。GC 含量是指鸟嘌呤和胞嘧啶含量的平均百分数，在真细菌中为 25%～75%。与细菌相比，真核生物的基因组更大也更为多变，但是他们的 GC 含量却非常一致（40%～45%左右）。每一个物种中，核苷酸组成都较为一致。

我们在图 12.6 中显示了 GC 含量的范围。

打碎的基因组 DNA 能够在离心后分成不同的浮力密度部分。一段同源 DNA 称为等容线，能够反映特征密度，也就是一个区域的 AT 或 GC 含量。基因一般都位于高 GC 含量的等容线区域。我们将在分析人类基因组时再讨论等容线。

二维细菌基因组显示（2DBGD）是一种电泳技术，和二维蛋白质胶相似，它也是先根据序列大小然后根据序列组成将细菌基因组 DNA 切断后分开。Malloff 等（2002）用 2DBGD 将 3 个高 GC 含量的呼吸病原体的基因组进行比较：*Bordetella pertussis*（68%GC），*M. tuberculosis*（66%GC）和 *Mycobacterium avium*（66%GC）这个技术也能在原核生物中检测插入和删除的发生。

14.8.2 寻找基因

细菌和古细菌具有较高的基因密度（大概每 1000 个碱基 1 个基因），没有内含子和很少有重复 DNA。这样比起在真核生物 DNA（第 15～第 17 章）中寻找基因就要相对简单一些了。现在有一些程序能够用来搜寻微生物基因（表 14-8）（Baytaluk 等，2002）。对于基因识别有用的基因组 DNA 有下面 4 个主要的特点。

表 14-8 在原核生物基因组中寻找基因的程序

程序	描述	网址
Critica	来自 Jonathan Badger, University of Waterloo	http://www.math.uwaterloo.ca/~jhbadger/critica/critica105/critica.html
ECOPARSE	邮件服务器,来自 Anders Krogh	http://www.cbs.dtu.dk/~krogh/Ecoparse.info
FrameD	定位基因和框移;优化 GC 含量高的基因组	http://genopole.toulouse.inra.fr/
GeneMark.hmm	使用隐马模型	http://intron.biology.gatech.edu/GeneMark
GeneMarkS	改进了基因起始位点预测的精度	http://dixie.biology.gatech.edu/GeneMark/genemarks.cgi
GLIMMER	在 TIGR	http://www.tigr.org/softlab/glimmer/glimmer.html
Orpheus	对细菌全基因组和大基因组片段的预测	http://pedant.gsf.de/orpheus

你能够在 L'Atelier BioInformatique de Marseille (http://www.up.univ-mrs.fr/~wabim/d_abim/rich-adn.html) 得到显示 GC 含量的在线程序或者 Emboss 程序 GEECEE (http://bioweb.paseur.fr/seqanal/interfaces/geecee.html)。还有许多其他显示 GC 含量的程序，如 GLIMMER（见下文）。

① 开放阅读框（ORF）长度。一个开放阅读框并不一定就是一个基因；例如，许多短开放阅读框并不是真实基因的一部分（下面将进一步讨论）。一个开放阅读框被一个起始密码子（如 ATG，编码甲硫氨酸）和一个终止密码子（TAA、TAG、TGA）所定义。然而，在细菌中还有一些可替换的起始密码子（如 GTG 或 TTG），但是很少有可替换的终止密码子。

② 在起始密码子的附近有结合核糖体的一致序列存在。有些情况下，可能在同一个阅读框中找到两个 ATG 密码子，每一个都有可能作为起始密码子。找到核糖体结合位点是判断起始密码子的重要方法。在细菌中，核糖体结合位点被称为 Shine-Dalgarno 序列。这是一段和 16S r RNA 的 3′端互补的多嘌呤核苷酸片段，从 −20（起始密码子的 5′方向）到 +13（起始密码子的 3′方向下游 13 个核苷酸）。Samuel Karlin 和他的同事们（Ma 等，2002）研究了 30 个原核生物基因组，得到了 Shine-Dalgarno 序列特征和基于密码子偏好性的基因表达水平、功能基因分类和起始密码子种类之间的关系。他们发现 Shine-Dalgarno 序列和高基因表达水平呈正相关。

③ 存在和基因一致的密码子使用模型。隐马尔可夫模型（第 10 章）在定义可能的蛋白质编码 DNA 序列时相当有用。

④ 推测的基因和其他已知基因有同源性。基因组 DNA 序列（包括推测的基因）能够在蛋白质数据库中用 blastx（参见第 5 章）进行查询。这个方法对在真核生物中查找基因有帮助。例如，外显子能够和表达序列标签（EST）相匹配（第 16 章）。

前 3 个特征被研究并被用作搜寻基因的基本特征。它们被称为基本特征是因为这些特征并不决定于其他物种的基因序列的比较。第 4 个特征（和其他基因的联系）被称为非基本特征。一些诸如 Critica（表 14-8）的原核生物基因搜寻程序同时结合了基本特征和非基本特征。

基本的方法有时候也成为从头开始的方法。

GLIMMER 系统是最重要的基因搜寻算法之一。它能够鉴别出细菌基因组中所有基因的 97%～98%（Delcher 等，1999a）。这个算法采用插入马尔可夫模型（IMMs）。一条马尔可夫链能够描述在一条基因组 DNA 序列上每一个核苷酸的概率分布（见第 10 章）。这个概率决定于序列上某个核苷酸之前的 k 个核苷酸。一条定阶的马尔可夫链能够描述每个核苷酸位置的 k 个碱基的

内容，例如，一条5阶的马尔可夫链模型描述了1024种可能分布，每种可能性都有5个核苷酸。k个碱基的片段集合被用作一个训练集来告诉算法哪种概率分布最可能和这个特殊基因组序列相关。k值大一点反映的信息更多。IMMs是马尔可夫模型的特殊形式，较少的k个核苷酸片段就被忽略了，常见的k个核苷酸片段序列有较大的权重。这个算法对每一个可能的编码区域进行了打分。

为了使用GLIMMER，必须在UNIX操作系统上运行这个程序。首先，为了训练算法，输入一个感兴趣的生物（如大肠杆菌）的基因组DNA数据集。这个的命令见图14-7(a)，对大肠杆菌76000个核苷酸的序列分析的输出结果见图14-7(b)。图中显示DNA片段的GC含量（这个例子中为51.5%）、参数（如最小基因长度设为90个核苷酸）、预测的基因列表包括基因编码方向（向前或者相反方向）、长度和打分。GLIMMER的输出还有对预测基因的总结，包括可能的基因重叠注释[图14-7(c)]。

GLIMMER是由TIGR的Owen White和他的同事们写的，GLIMMER是Gene Locator and Interpolated Markov Modeler的首字母缩写词。

对原核基因的预测也有一些困难。

① 一个基因组DNA片段可能有多个基因在相同方向或者相反方向的可选阅读框中。

② 很难判断一个短开放阅读框是否真正被转录了。根据Skovgaard等（2001）的研究结果，在很多基因组中有太多的短基因被注释了。对于大肠杆菌，他们认为只有3800个真正的基因，虽然现在有4300个基因被注释了。因为终止密码子（TAA、TAG、TGA）AT含量比较高，GC含量高的基因组倾向于含有较少终止密码子并有更多的预测较长的开放阅读框。对于所有基因组中预测的蛋白质，假定的蛋白质（定义为没有经过实验验证它们是否表达的蛋白质）的比例随着序列长度的缩短也升高。

③ 一个开放阅读框编码一个基因，并有一个终止密码子。如果基因组DNA中同一条链上发生连续编码其他阅读框，就可能发生阅读框移动。可能发生阅读框移动是因为排列错误或者发生突变产生假基因（无功能基因）。

④ 有一些基因属于操纵子，经常在原核生物中起着相关作用。操纵子有启动子和终止子序列模体，但是对这些并没有很好地研究过。Steven Salzberg和他的同事们（Ermolaeva等，2001）分析了34个细菌和古细菌基因组中很可能属于同一种操纵子的7600对基因。

一个操纵子是一簇从同一个启动子开始转录、引起多顺反子的基因。这个研究预测的基因对（包括73个细菌和古细菌的基因组）能够在网站 http://www.tigr.org/tigr-scripts/operons/operons.cig 上得到。

⑤ 细菌和古细菌中经常会发生水平基因转移。我们接下来就讨论这个问题。

14.8.3 水平基因转移

水平基因转移（LGT）是一个基因组从另外一个物种得到某个基因而非从祖先遗传得到的现象（Eisen，2000；Koonin等，2001）。很可能在一个生物中鉴别出另一个非常远源的生物的同源蛋白。这个物种如何得到这个基因最简单的解释就是发生了水平基因转移。这个机制是基因组进化过程中一个很大的动力。基因转移是单方向的而不是DNA的相互交换。

水平基因转移现象非常重要，有以下几个原因。

① 这个机制和正常的遗传模式有很大的不同，正常遗传是基因从亲代传给子代。这样，水平基因转移改变了进化的概念。

(a)

```
$ /usr/local/glimmer2.02/build-icm < ecoli_first100.txt > trainecoli

$ /usr/local/glimmer2.02/glimmer2 ecoli76k.fasta colitrain
```

(b)

```
GC Proportion = 51.5%
Minimum gene length = 90
Minimum overlap length = 30
Minimum overlap percent = 10.0%
Use independent scores = True
Ignore independent score on orfs longer than 765
Use first start codon = True
```

ID#	Fr	Orf Start	Gene Start	End	Lengths Orf	Lengths Gene	Gene Score	-- Frame Scores - F1	F2	F3	R1	R2	R3	Indep Score	
	F2	35	44	178	144	135	0	0	0	0	_	0	_	99	0
	F1	226	247	402	177	156	0	0	_	_	_	0	_	99	0
	F3	273	420	563	291	144	0	_	_	0	_	_	0	99	0
	R2	740	713	609	132	105	0	_	6	_	_	0	2	89	0
	F2	515	548	916	402	369	0	_	0	_	_	_	0	99	6
	R1	1149	1143	1036	114	108	0	_	_	0	0	_	0	99	0
	F3	888	936	1265	378	330	0	_	_	0	_	_	0	99	0
	F1	1162	1210	1347	186	138	0	0	_	_	_	_	0	99	0
	F3	1365	1377	1592	228	216	0	_	_	0	_	_	0	99	0
	F3	1707	1710	1823	117	114	0	_	_	0	_	_	0	99	0
1	R3	1951	1909	380	1572	1530	99	_	_	_	_	_	99	0	101
	F3	1857	1872	1994	138	123	0	_	_	0	_	_	_	99	0
	R1	2124	2121	2029	96	93	0	_	_	_	0	_	_	99	0
	F3	2043	2043	2273	231	231	0	_	_	0	_	_	0	99	0
	F1	2098	2140	2319	222	180	0	0	0	_	_	_	0	99	0
	R1	2604	2589	2182	423	408	0	_	0	_	0	_	0	99	0
	F3	2313	2349	2645	333	297	0	_	0	0	_	_	0	99	0
	F2	2057	2183	2692	636	510	0	_	0	_	_	_	0	99	0
2	R1	2844	2835	2692	153	144	90	_	_	_	90	_	1	7	90
	F1	2809	2815	2946	138	132	7	7	_	0	_	_	4	87	7
	F3	2769	2835	2987	219	153	0	_	_	0	_	_	5	94	0
3	R3	3034	2971	2039	996	933	99	_	_	_	_	_	99	0	109

(c)

```
48   66485   65460   [-3 L=1026]
49   69967   68642   [-2 L=1326]
50   71402   70614   [-3 L= 789]
51   73146   71341   [-1 L=1806]
52   73426   74445   [+1 L=1020]
53   74625   76166   [+3 L=1542]
54   80198   76878   [-3 L=3321]
55   80118   81140   [+3 L=1023]   [LowScoreBy #56 L=918 S=2]
56   81188   80223   [-3 L= 966]   [OlapWith #55 L=918 S=97]
57   82370   81288   [-3 L=1083]
58   84895   83759   [-2 L=1137]   [DelayedBy #59 L=108]
59   84873   86438   [+3 L=1566]
```

图 14-7 GLIMMER 程序（TIGR，http://www.tigr.org）在细菌 DNA 中寻找新基因相当有效。这个程序只能在 UNIX 操作系统上运行。(a) 首先要训练一个基因组 DNA 的数据集来产生一个马尔科夫模型，然后才能运行程序。(b) 输出结果中包括前面或者相反的链上的开放阅读框数据，还有找到基因的可能性分值。(c) 结果中还有一些基因的列表。这里用到的 DNA 来自大肠杆菌 K12（U14003）

② 这个机制在原核生物中相当普遍，真核生物中也发现了一些例子。在生命的 3 个主要分支之间和之内都观察到了这种机制，但是在细菌中最为普遍。

③ 水平基因转移能够很大程度上混淆系统发生分析。如果一个被选来进行系统发生分析 DNA、RNA 或者蛋白质发生了水平基因转移，那么产生的系统发生树就不能正确反映研究的物种的进化情况了。对水平基因转移的一个极端的解释是如果水平基因转移非常普遍的话，那么原则上就不可能构建出一个正确的发生树。

水平基因转移是一个多步骤过程（图 14-8）（Eisen，2000）。一个家族中的基因（传统的达尔文遗传过程）可能转移到另外一个家族中去。这种 DNA 转移能够被病毒载体或者像同源重组这样的机制所调节。一旦一个新基因被整合到一个种群个体的基因组中（如图 14-8 中的物种 3），这些个体中的正选择保留了这些基因。为了能够在新物种种群中被保留、繁殖和扩散，一个转移基因需要给新物种带来好处。重组后新基因就适应了这个新家族，这个过程称为“改良”（Eisen，2000）（图 14-8，第 6 个箭头）。

Carl Woese（2002）认为在早期的进化过程中水平基因转移是主流，最初的细胞进化是一个普遍的过程，后来才是垂直（达尔文）进化。

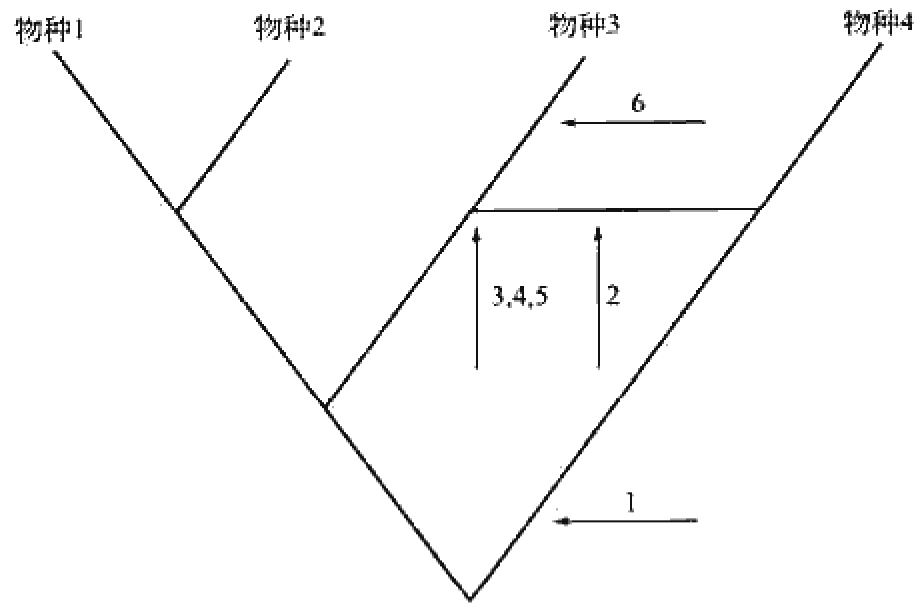


图 14-8 水平基因转移在不同的阶段发生。在这个假设中，4 个物种从一个共同的祖先进化而来。每个物种随着时间的流逝多少遗传一些祖先的基因（箭头 1）。在接下来的某一点上，一个基因从物种 3 水平转移到了物种 4（箭头 2）。转移的基因必须固定在一个基因组中（箭头 3），在正选择下依然保留（箭头 4），并且在物种 3 的种群中传播（箭头 5）。水平转移基因然后进化为一个新基因组的独立部分（箭头 6）。这个基因和物种 3 中的其他基因有很大的区别，其核苷酸组成和密码子使用偏好都和物种 4 中的基因相似。本图根据 Eisen (2000) 改编，授权使用

作为一个特殊的水平基因转移例子，我们来研究一下人类蛋白质胸腺嘧啶核苷磷酸化酶（P19971）。Salzberg 等（2001）找出了一些从细菌中水平转移到人类的基因，这个蛋白质是由其中之一编码的。在很多数据库中用 Blastp 查找这个蛋白，找到了很多低期望值（高分）的匹配。根据分类结果，这个蛋白最好的匹配就是它自己（图 14-9，箭头 1）、大鼠和小鼠的同源蛋白（箭头 2 和箭头 3）。其他数据库中匹配的都是如大肠杆菌和炭疽杆菌之类的原核生物中的蛋白。因此这个基因可能就是通过水平基因转移进人人类基因组的。根据 2001 年获得的人类基因组序列草图（国际人类基因组测序中心，2001；Ponting，2001），认为人类基因组中有几百个基因是水平基因转移得到的，随后的分析认为实际数量远小于这个数目（Stanhope 等，2001；Salzberg 等，2001）。

怎样判别水平基因转移呢？最主要的标准就是一个基因有不同寻常的核苷酸组成、密码子偏好、系统发生位置或者其他特征来与同一个基因组中的其他基因相区别。有如下 3 个最主要的原则来判断是否为水平基因转移：

```
Organism Report
  Homo sapiens (man) [mammals] taxid 9606
 sp|P19971|TYPH_HUMAN Thymidine phosphorylase precursor (Td...    941  0.0 ← 1
 gb|AAA60043.1| endothelial cell growth factor >gi|6862560|...    938  0.0
 gb|AAH18160.1|AAH18160 endothelial cell growth factor 1 (p...    937  0.0

  Mus musculus (mouse) [mammals] taxid 10090
 ref|NP_612175.1| endothelial cell growth factor 1: thymidi...    732  0.0 ← 2

  Rattus norvegicus (brown rat, ...) [mammals] taxid 10116
 ref|XP_235562.1| similar to thymidine phosphorylase [Mus m...    695  0.0 ← 3

  Deinococcus radiodurans [eubacteria] taxid 1299
 pir||A75520 pyrimidine-nucleoside phosphorylase - Deinococ...    322  1e-86
 ref|NP_295490.1| anthranilate synthase component II [Deino...     62  2e-08
 sp|Q9RTJ5|TRPD_DEIRA Anthranilate phosphoribosyltransferase       60  4e-08

  Mycobacterium tuberculosis CDC1551 [high GC Gram+] taxid 83331
 ref|NP_337943.1| thymidine phosphorylase [Mycobacterium tu...    306  4e-82

  Geobacillus stearothermophilus [eubacteria] taxid 1422
 pdb|1BRW|A Chain A, The Crystal Structure Of Pyrimidine Nu...    299  7e-80
 pir||JT0875 pyrimidine-nucleoside phosphorylase (EC 2.4.2....    275  2e-72
 gb|AAD33792.1|AF139534_3 anthranilate phophoribosyl transf...     51  4e-05

  Sinorhizobium meliloti [a-proteobacteria] taxid 382
 ref|NP_384236.1| PROBABLE THYMIDINE PHOSPHORYLASE TRANSMEM...    298  2e-79
 sp|Q92PS0|TRPD_RHIME Anthranilate phosphoribosyltransferas...     51  3e-05

  Thermoanaerobacter tengcongensis [eubacteria] taxid 119072
 ref|NP_622143.1| Thymidine phosphorylase [Thermoanaerobact...    296  5e-79
 sp|Q8R9M6|TRPD_THETN Anthranilate phosphoribosyltransferas...     52  2e-05

  Streptomyces coelicolor A3(2) [high GC Gram+] taxid 100226
 ref|NP_629043.1| thymidine phosphorylase [Streptomyces coe...    295  1e-78

  Salmonella enterica subsp. enterica serovar Typhi [enterobacteria] taxid 90370
 pir||AD1073 thymidine phosphorylase (EC 2.4.2.4) [imported...    286  4e-76
 pir||AE0653 anthranilate synthase component II, anthranila...     51  4e-05

  Escherichia coli CFT073 [enterobacteria] taxid 199310
 ref|NP_757312.1| Thymidine phosphorylase [Escherichia coli...    284  2e-75
 gb|AAN80194.1|AE016760_53 Anthranilate synthase component ...     52  2e-05

  Escherichia coli O157:H7 EDL933 [enterobacteria] taxid 155864
 ref|NP_290996.1| thymidine phosphorylase [Escherichia coli...    284  2e-75
```

图 14-9 用人类 P19971（胸苷磷酸化酶）进行 tblastp 搜索得到的分类报告。这个基因被认为可能是从细菌到人类的水平转移基因。匹配的数据库包含人类（箭头 1）、小鼠（箭头 2）、大鼠（箭头 3）和一些其他细菌。没有其他真核生物的数据库和这个查询匹配

> 你能够在 Salzaberg 等（2001）的补充网站（http://www.sciencemag.org/cgi/data/1061036/DC1/1）上看到几十个已经被整合到人类基因组中的基因列表。

① 比较不同基因的系统发生树，这是最好的一个方法（Eisen，2000）；

② 可以利用基因组中每一个基因最好的匹配模式；

③ 能够通过评估物种间的基因分布模式来搜寻经过水平基因转移的基因。例如，如果一个基因只存在于 Crenarcheota 和植物中，但不存在于其他古细菌、细菌或者真核生物中，这个就可以作为证据表明发生了水平基因转移机制。如果只是发生了基因丢失或者快速突变而没有发生水平基因转移，这种方法可能产生人为的结果（假阳性）。

让我们重新来讨论一下胸腺嘧啶核苷磷酸化酶。为了确定是否发生水平基因转移有以下 3 个理由要考虑。

① 如果该基因存在于像果蝇这样的昆虫或者植物中，那么对于水平基因转移的争辩就会减少很多。黑猩猩中有胸腺嘧啶核苷磷酸化酶吗？除非其基因组完全被测序了（或者我们直接克隆出黑猩猩基因组中的每一个基因）否则我们不知道答案。另外一个用来判断是否发

生了水平基因转移的重要证据就是看这个蛋白质是否存在于整个生命发生树上，但是我们可能没有足够的其他物种的数据。在 EST 数据库中用 tblastn 查找胸腺嘧啶核苷磷酸化酶时发现确实有很多真核生物有这个基因（表 14-9）。这样，从这个角度来看这个基因可能没有发生水平基因转移。在细菌和古细菌基因组中，我们才开始研究全基因组多样性，逐渐地会更容易理解进化关系。

② 也可能考虑的基因发生了快速突变，这样系统发生信号就丢失了。

③ 另外一种可能性是曾经在其他真核生物物种中存在的基因后来丢失了。

表 14-9 在 EST 数据库中用人类胸苷磷酸酶作为输入进行 tblastn 搜索后的结果（2002 年 10 月）

生物体	俗名	编号	期望值
Salmo salar	大西洋鲑	BG935885	2×10^{-72}
Oncorhynchus mykiss	虹鳟鱼	CA348293	3×10^{-49}
Gallus gallus	鸡	BM491333	1×10^{-56}
Tetrahymena thermophila	四膜虫	BM400607	8×10^{-47}
Drosophila melanogaster	果蝇	BI591052	3×10^{-27}
Rattus norvegicus	老鼠	AW914201	2×10^{-22}

14.8.4 注释和比较

随着原核生物基因组不断被测序，它们也在不断被注释（见第 12 章）。在细菌和古细菌中这个过程比起真核生物更为简单。在 TIGR CMR、NCBI 和 EBI 中有很多可用工具。我们先从 NCBI 开始。同源基因分类（COG）数据库收集了包括啤酒酵母（第 15 章）在内的许多原核生物基因组信息（Tatusov 等，1997，2001）。COG 的目标就是提供一个原核生物蛋白质的系统发生分类。其方法就是根据 BLAST 搜索结果的最好匹配将蛋白质分类。

COG 主页如图 14-10 所示。COGnitor 链接允许输入一段感兴趣的序列在 COG 数据库中进行查找。可以进行文本查询，输入“lipocalin”后就链接到了一组原核生物的 lipocalin（分类号为 COG3040）。图 14-11(a) 是一组不同种的细菌 lipocalin 之间的联系，图 14-11(b) 是这些蛋白质之间的多序列比对。注释“M”[图 14-11(a)] 表示每一类同源基因的功能分类（M 表示外膜蛋白，如细菌 lipocalins）。

> COG 链接是 http://www.ncbi.nlm.nih.gov/COG/。现在大约有超过 100000 个蛋白质被分到 75000 个同源基因簇中。

从主页（图 14-10）能够链接到 COG 功能分类（图 14-12），这个分类系统和基因分类系统（第 8 章）有点相似，这里，主要的功能种类有信息储存和处理、细胞过程和代谢作用。

直系同源基因类的分布可以用来有效地鉴别一组出现频率很高或很低的相关蛋白质。例如，在 26 个不同的物种中发现了 84 个直系同源基因簇（图 14-13，底部）。这些相当保守的蛋白质家族包括 tRNA 合成酶、核糖体蛋白质和其他如信号识别粒子 GTPase 和甲基转移酶之类的酶。这样，它们中的每一个都可以用来进行系统发生分析。还有一个极端的例子，即在 COG 数据库中三个物种之间就有 493 个直系同源基因分类（图 14-13，顶部）。这些记录中包括许多注释为“未研究”或者“预测”的蛋白质（如“未研究的膜蛋白”）。如果你正在研究一个特殊的原核蛋白质，那就可能有兴趣来研究这些蛋白质，因为它们可能在那个物种中相对比较独特。

> 我们在表 8-9 中列出了 COG 的功能分类。

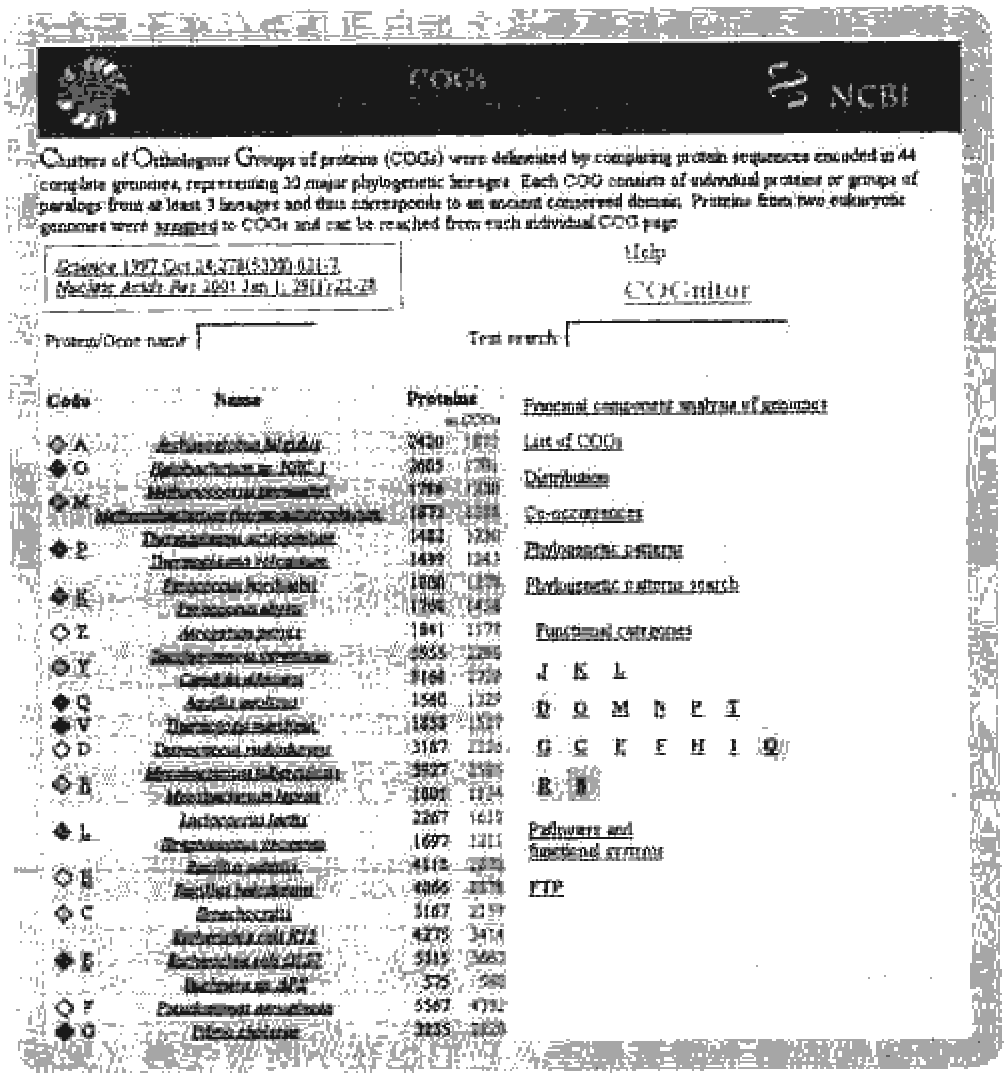
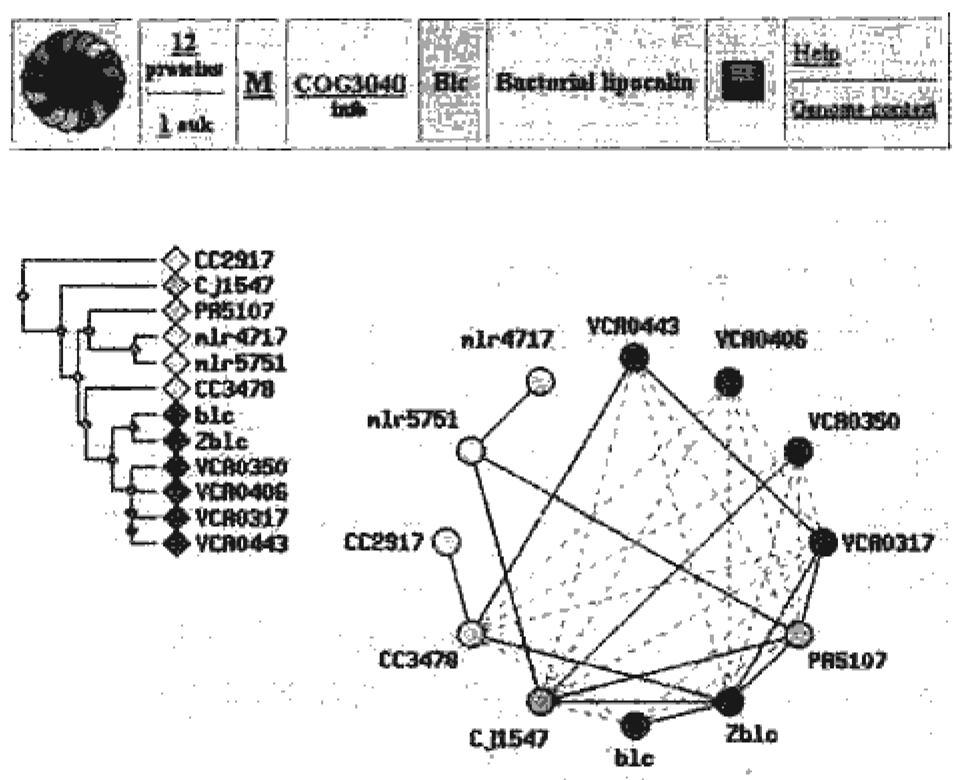


图 14-10 NCBI 网站上的 COG 页面提供了与已经完成测序的基因组功能上相关的基因和蛋白质分析（http://www.ncbi.nlm.nih.gov/COG/）

(a)



(b)

```
CLUSTAL W (1.74) multiple sequence alignment


blc          MRLLPLVAAATAAF--LVVACSSP--TPPRGVTVVNNFDAKRYLGTWYEIARFDHRFERG
Zblc         MRLLPLVAAATAAF--LVVACSSP--TPPRGVTVVNNFDAKRYLGTWYEIARFDHRFERG
VCA0317      MRAIFLILCA---V--LLNGCLG----NPESVKPVSDFELNNYLGKWYEVARLDHSFERG
VCA0443      MRAIFLILCS---V--LLNGCLG----NPESVKPVSDFELNNYLGKWYEVARLDHSFERG
VCA0406      MRAIFLILCS---V--LLNGCLG----NPESVKPVPDFELNNYLGKWYEVARLDHSFERG
VCA0350      MRAIFLILCS---V--LLNGCLG----NPESVKPVSDFELNNYLGKWYEVARLDHSFKRG
mlr4717      MFGGWIDRKE---------NFNTD----TAAVTAITS-LNLSRYLGKWYELCRLPIKYEDE
mlr5751      MFRGWIDGKE---------NFNTD----TAAVTAITS-LNLSQYLGKWYELCRLPIKYEDE
PA5107       MIGLPIRVLCTLWLGLLLVGCAG----SFKNPPATEQVDLQRYQGTWYELARLPMFFQRN
CC3478       MRKSLLALPVILSLTTTLGACASGPVGNPNSPQPAKAVELDRYLGKWYEVARYDHRFQRG
CC2917       MLNTAIALFA--------LAQAAAV---QPQTPLIDP---ARHYSGVWNEVGRTPHRITKG
Cj1547       ---------------------------------MLKLIGELSLQNYNGKWLKIARKPAFFQKS
                                                  .* * * *: *
```

图 14-11 （a）在 COG 页面上选择“细菌 lipocalin 蛋白”的链接，就出现 12 个细菌蛋白质和 1 个真核生物蛋白质（来自果蝇）的页面。12 个蛋白质用两种分类树显示。（b）点击 12 个蛋白质的链接，就会出现由 ClustalW 产生的多序列比对

Code	COGs	Domains	Description	Pathways and functional systems
Information storage and processing				
J	217	6449	Translation, ribosomal structure and biogenesis	4
K	132	5438	Transcription	3
L	184	5337	DNA replication, recombination and repair	2
Cellular processes				
D	32	842	Cell division and chromosome partitioning	-
O	110	3165	Posttranslational modification, protein turnover, chaperones	-
M	155	4079	Cell envelope biogenesis, outer membrane	1
N	133	3110	Cell motility and secretion	2
P	160	5112	Inorganic ion transport and metabolism	1
T	97	3627	Signal transduction mechanisms	-
Metabolism				
C	234	5594	Energy production and conversion	7
G	171	5262	Carbohydrate transport and metabolism	4
E	233	8383	Amino acid transport and metabolism	10
F	85	2364	Nucleotide transport and metabolism	5
H	154	4057	Coenzyme metabolism	11
I	75	2609	Lipid metabolism	2
Q	62	2754	Secondary metabolites biosynthesis, transport and catabolism	-
Poorly characterized				
R	449	11948	General function prediction only	-
[illegible]	[illegible]	[illegible]	Function unknown	[illegible]

图 14-12 COG 数据库一个主要的特征就是将蛋白质注释为 18 类。这种注释对于决定在细胞中起作用的生化途径非常重要

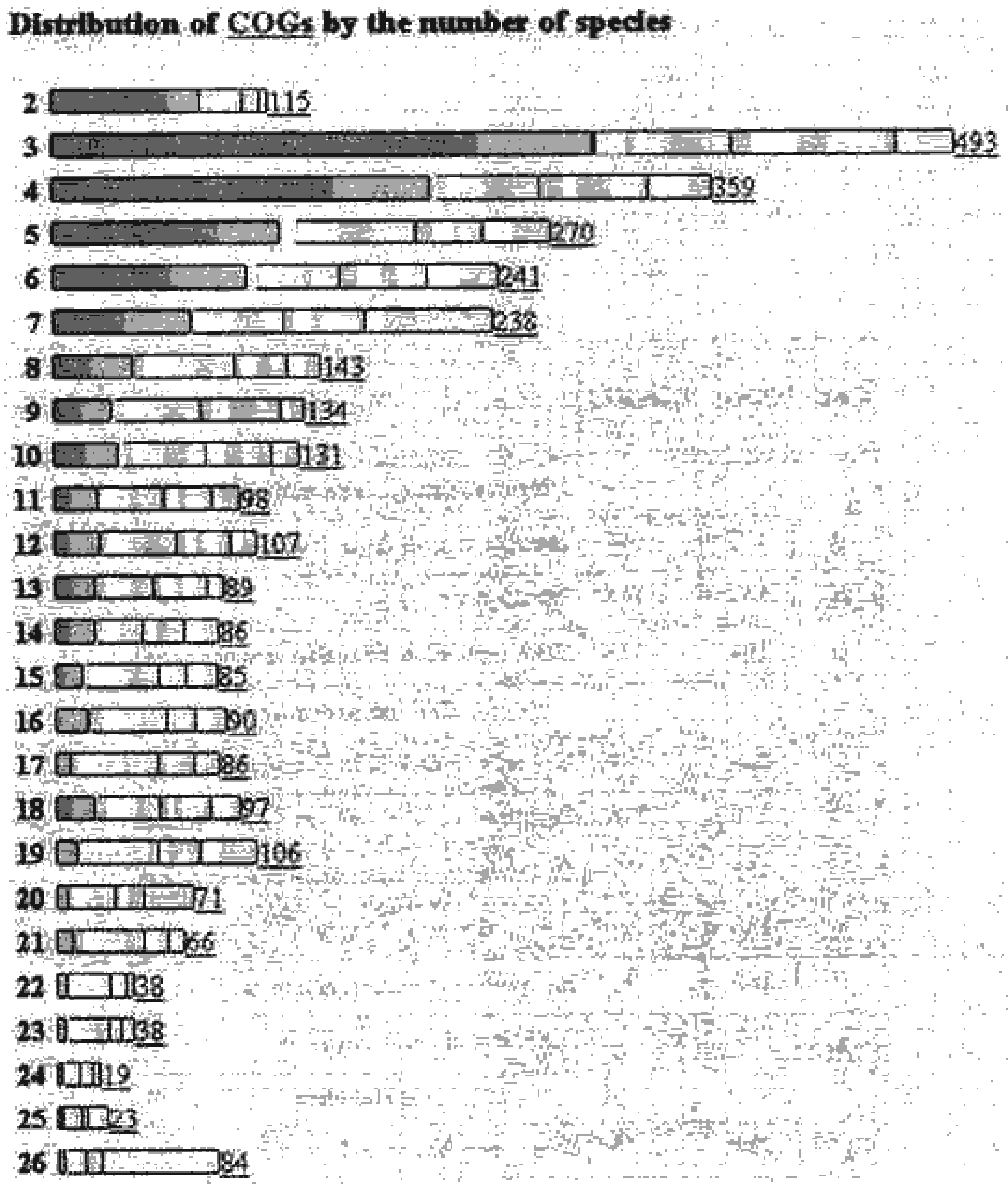


图 14-13 COG数据库中的分布表明大部分基因簇集中在3～7个物种中。在COG数据库中26个物种共有84个同源基因编码的蛋白类，包括tRNA合成酶、核糖体蛋白质、激酶和其他在地球上的生命中相当保守的蛋白质

14.9 原核生物基因组的比较

14.9.1 COG

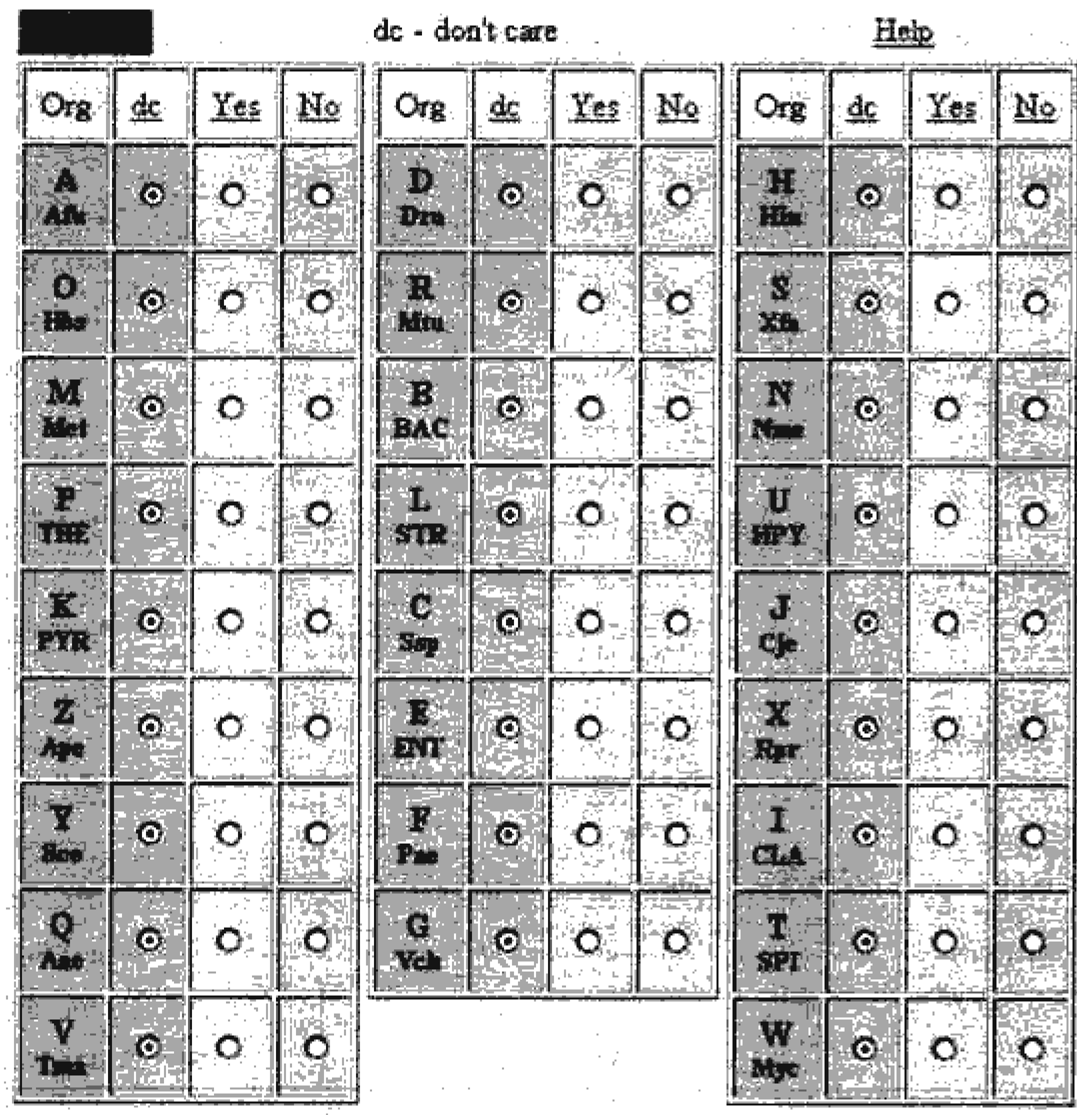
在全基因组测序过程中最重要的收获之一就是比较分析能够帮助我们更好地理解基因组。比较近源或者远源基因组对我们都有帮助。表14-10显示的是一些已经完成测序的近源基因组生物。比较这些基因组很有意义，有以下几个原因：

① 我们能够发现为什么有些菌株是致病的；

② 最后，我们能够根据病原体的基因型来预测感染后的临床结果；

③ 我们可以发明一些接种疫苗的方法。

继续我们对COG的讨论，这个数据库的特点之一是系统发生模式搜索（图14-10），能够比较各个物种。在搜索页面上（图14-14）选择 *Chlamydia trachomatis* 和 *Chlamydia pneumoniae*。*Chlamydiae* 是必需胞内细菌，在系统发生树上和其他细菌类别离得比较远。*Chlamydia pneumoniae* 感染人类，导致肺炎和支气管炎。*Chlamydia trachomatis* 导致沙眼（一种眼疾，可能会导致失明）和性传播疾病。为什么这两个近源的细菌会感染不同的身体区域并且产生不同的病症呢？对它们基因组的测序工作已经完成，并进行了比较（Stephens 等，1998；Kalman 等，1999；Read 等，2000）。



Differences in closely related genomes

J ALPHA	*Mesorhizobium loti*	vs.	*Caulobacter crescentus*
L STR	*Lactococcus lactis*	vs.	*Streptococcus pyogenes*
P THE	*Thermoplasma acidophilum*	vs.	*Thermoplasma volcanium*
H PAS	*Haemophilus influenzae*	vs.	*Pasteurella multocida*
M MET	*Methanococcus jannaschii*	vs.	*Methanobacterium thermoautotrophicum*
R MYb	*Mycobacterium tuberculosis*	vs.	*Mycobacterium leprae*
N NEI	*Neisseria meningitidis MC58*	vs.	*Neisseria meningitidis Z2491*
K PYR	*Pyrococcus horikoshii*	vs.	*Pyrococcus abyssi*
B BAC	*Bacillus subtilis*	vs.	*Bacillus halodurans*
E ENT	*Escherichia coli*	vs.	*Buchnera sp. APS*
U HPY	*Helicobacter pylori 26695*	vs.	*Helicobacter pylori J99*
I CLA	*Chlamydia trachomatis*	vs.	*Chlamydia pneumoniae*
W MYC	*Mycoplasmas*	vs.	*Ureaplasma urealyticum*
T SPI	*Treponema pallidum*	vs.	*Borrelia burgdorferi*

图 14-14　在 COG 站点上允许进行特殊基因组中的同源基因比较。这种分析能够帮助研究为什么一种细菌是致病的，而另外一种非常近源的却不致病，或者为什么古细菌能够适应某一种生态环境

表 14-10 已经完成测序的一些近源基因组原核生物

生物体	编号	基因组大小/bp
Chlamydophila pneumonide AR39	NC_002179	1229858
C. pneumoniae CWL029	NC_000922	1230230
C. pneumoniae J138	NC_002491	1226565
Escherichia coli K12	NC_000913	4639221
E. coli O157 : H7	NC_002695	5498450
E. coli O157 : H7 EDL933	NC_002655	5528445
Helicobacter pylori 26695	NC_000915	1667867
H. pylori J99	NC_000921	1643831
Mycobacterium tuberculosis CDC1551	NC_002755	4403836
M. tuberculosis H37Rv	NC_000962	4411529
Neisseria meningitidis MC58	NC_003112	2272351
N. meningitidis Z2491	NC_003116	2184406
Staphylococcus aureus aureus MW2	NC_003923	2820462
S. aureus aureus Mu50	NC_002758	2878040
S. aureus aureus N315	NC_002745	2813641
Streptococcus agalactiae 2603V/R	NC_004116	2160267
S. agalactiae NEM316	NC_004368	2211485
Streptococcus pneumoniae R6	NC_003098	2038615
S. pneumoniae TIGR4	NC_003028	2160837
Streptococcus pyogenes M1 GAS	NC_002737	1852441
S. pyogenes MGAS315	NC_004070	1900521
S. pyogenes MGAS8232	NC_003485	1895017

在美国，10%的肺炎和5%的支气管炎是由 *C. pneumoniae* 引起的。

我们能够从 COG 搜索工具中比较 *C. pneumoniae* 已经注释功能的蛋白质和 *C. trachomatis* 已经注释功能的蛋白质（图 14-15）。其中显示了两种生物中共有的同源蛋白质和各自所特有的蛋白质。有几百个蛋白质是各个细菌所特有的，包括外膜蛋白质家族，这个蛋白质家族对于组织定向相当重要（Kalman 等，1999）。

14.9.2 TaxPlot

NCBI 提供了一个非常有效的基因组比较工具，并且使用起来非常容易。从基因组查询页面上，选择 *C. trachomatis* 就得了如图 12-9 所示的页面。选择 TaxPlot，你就能够将两个基因组和一个参考基因组（图 14-16 中的例子是炭疽病菌 *B. anthracis*）进行比较（如 *C. trachomatis* 和 *C. pneumoniae* AR39 的基因组）。在这个图上，每一个点都代表参考基因组中的一个蛋白质。x 坐标和 y 坐标显示了两个被比较的 *Chlamydia* 蛋白质组中每个蛋白质最佳匹配的 BLAST 分值。大多数蛋白质都在图的对角线上，表明它们在参考蛋白和 *Chlamydia* 蛋白中的分值相同（或者几乎相同）。然而，也有值得注意的异常值，代表了两种生物不同表型的重要基因。这些点是可以点击的（参见图 14-16 中带圆圈的数据点），选择的数据点是高亮的（图 14-16，箭头 4）。这个蛋白质被鉴别出来为一个 ABC 转运蛋白，还有一些链接可以链接到成对的 BLAST 比较（图 14-16，箭头 5）。TaxPlot 还能根据 COG 分类系统规则在图上标上颜色。

Chalmydia trachomatis	-	x	-	x	
Chalmydia pneumoniae	-	-	x	x	
Function	2574	8	22	562	
J K L	328	2	3	185	**Information storage and processing**
J	99	-	2	112	Translation, ribosomal structure and biogenesis
K	100	1	-	23	Transcription
L	129	1	1	50	DNA replication, recombination and repair
D O M N P T	500	-	2	138	**Cellular processes**
D	24	-	-	7	Cell division and chromosome partitioning
O	81	-	-	25	Posttranslational modification, protein turnover, chaperones
M	113	-	1	34	Cell envelope biogenesis, outer membrane
N	85	-	-	36	Cell motility and secretion
P	129	-	1	22	Inorganic ion transport and metabolism
T	68	-	-	14	Signal transduction mechanisms
C G E F H I Q	770	5	14	189	**Metabolism**
C	184	1	-	35	Energy production and conversion
G	126	-	-	31	Carbohydrate transport and metabolism
E	179	3	5	40	Amino acid transport and metabolism
F	64	-	5	16	Nucleotide transport and metabolism
H	113	-	4	35	Coenzyme metabolism
I	48	1	-	26	Lipid metabolism
Q	56	-	-	6	Secondary metabolites biosynthesis, transport and catabolism
R S	1045	1	5	86	**Poorly characterized**
R	367	1	4	60	General function prediction only
S	678	-	1	26	Function unknown

图 14-15 使用 COG 系统发生模式搜索工具比较 *C. trachomatis* 和 *C. pneumoniae* 的同源蛋白。*C. pneumoniae* 导致了美国 10% 的肺炎和 5% 的支气管炎，而 *C. trachomatis* 导致沙眼（一种眼睛疾病，可能会导致失明）和性传播疾病。对由这两个基因组编码的蛋白质进行比较就能够解释它们导致感染的不同模式。这个 COG 分析表明许多蛋白质只在某些细菌中存在

因此 TaxPlot 是一种在两种感兴趣的不同微生物中找出不同蛋白的简单工具。这个工具也被运用到真核生物中（第 16 章）。

14.9.3 MUMmer

在整个微生物基因组的比对中最大的挑战就是用动态程序（第 3 章）比对上百万的碱基对所需要的大量时间。我们在第 5 章介绍了一些如 BLAST 这样的快速算法。然而对于基因组比对来说，这些工具还比较初级（Miller，2001）。MUMmer 软件包提供了一个对微生物基因组进行快速准确比对的方法（Delcher 等，1999b）。最近，经过对算法改进后，它也能够对真核生物序列进行比对了（Delcher 等，2002）。

MUMmer 是由 TIGR 的 Steven Salzberg 和他的同事们写的。你能够在 http://www.tigr.org/software/mummer/ 上得到它。还有一个交互的网页（http://www.tigr.org/tigr-scripts/CMR2/webmum/mumplot）。

MUMmer 将两条序列作为输入。图 14-17 显示的是 MUMmer 网页浏览器。这个算法找到了所有的长于一个设定的最小长度值 k 并且很好匹配的子序列。根据定义，这些匹配序列是最小的，因为如果将它们向任意方向延长一点就会导致不匹配。这个算法采用后缀树，这是一种在成对比对中找到最大惟一匹配（MUM）的搜索结构的算法。MUMs 是有序的，并且这个算法通过找出大插入片段、重复片段、小的突变区域和单核苷酸多样性（SNPs）序列来除去间隔。

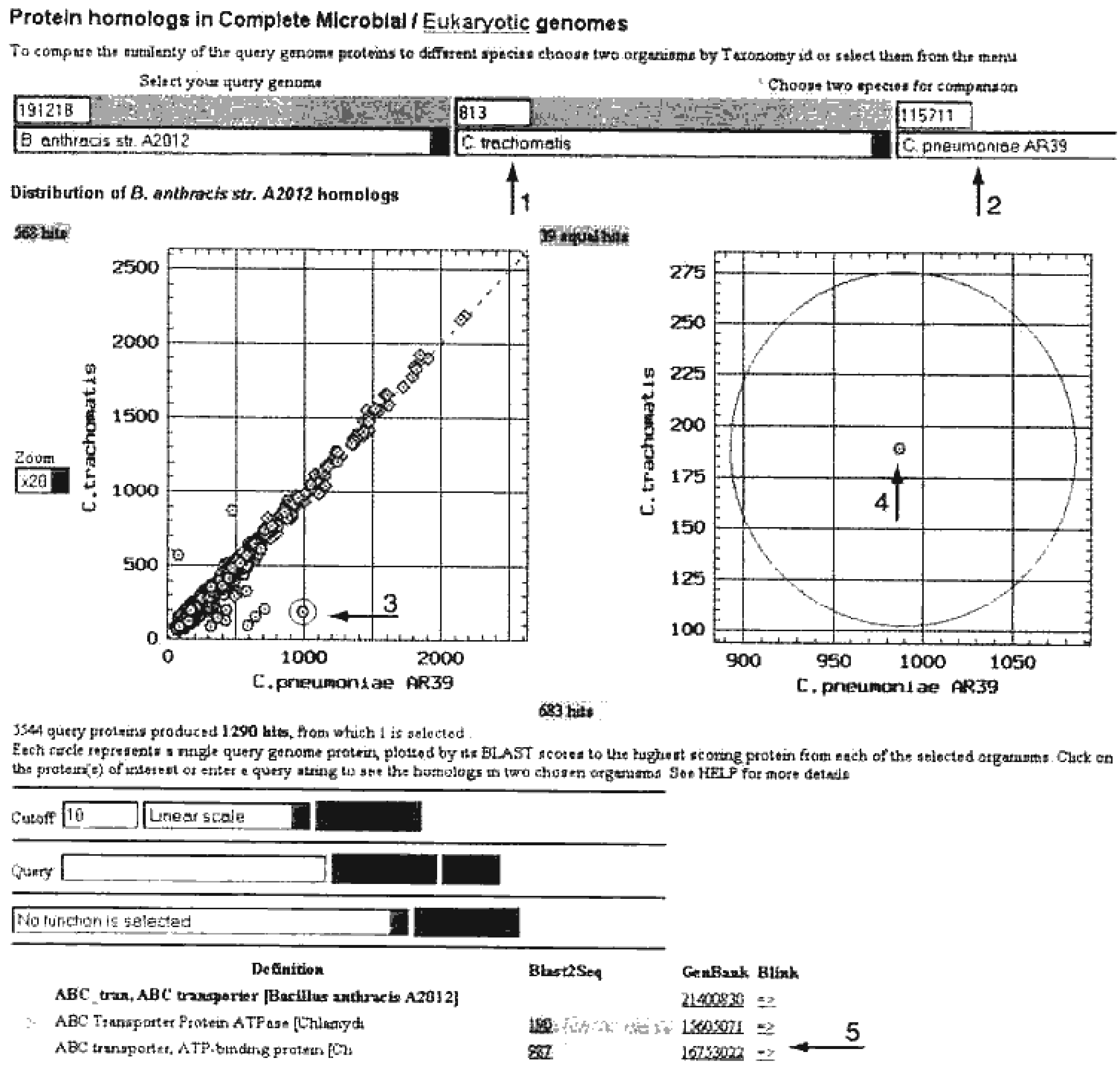


图 14-16 NCBI 上的 TaxPlot 工具能够将两个细菌基因组（*C. trachomatis* 和 *C. pneumoniae* AR39）和一个参考基因组进行比较（*B. anthracis* A2012）。每个点表示每个细菌相对于参考基因组的 BLASTP 分值的分布。一共找到了 39 个匹配，大约有 600 个匹配至少和 *C. trachomatis* 或 *C. pneumoniae* 有微弱的关系

> 当我们只注意到 TIGR 和 NCBI 上的工具的时候，欧洲生物信息学研究所也提供了一些重要的微生物资源。ARTEMIS（Mural，2000）是另外一个对微生物基因组比对和分析非常有用的工具（http://www.sanger.ac.uk/Software/Artemis/）。

MUMmer 的输出结果由点阵图组成（图 14-18），该结果以最小比对长度（如 15 个或 100 个碱基对）为序，显示了两个基因组序列的比对结果。结果包括如下内容：

① SNPs；

② 比单个 SNP 更加分散的序列区域；

③ 大的插入片段（例如，经过转座、序列逆转和水平基因转移）；

④ 重复片段（例如，一个基因组中的复制）；

⑤ 前后串联复制（不同的拷贝数）。

在图 14-18 中的例子中，将两个大肠杆菌菌株进行比较：无毒的大肠杆菌 K12 和大肠杆菌 O157：H7（在污染食物中有这个菌

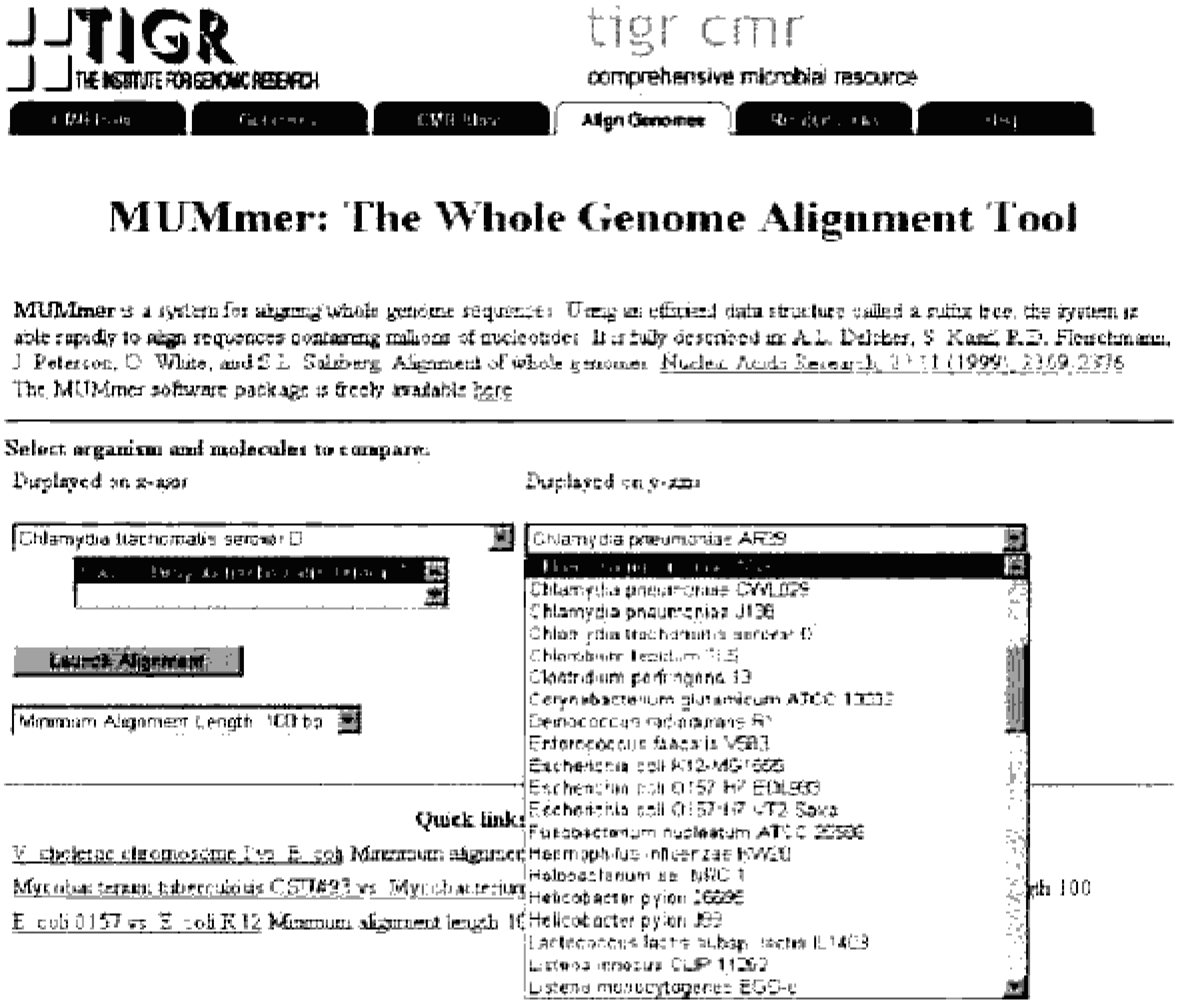


图 14-17 TIGR 上的 MUMmer 网页浏览器能够允许选择两个感兴趣的微生物基因组在点阵图上进行比较。比对的最短序列能够进行自我调整

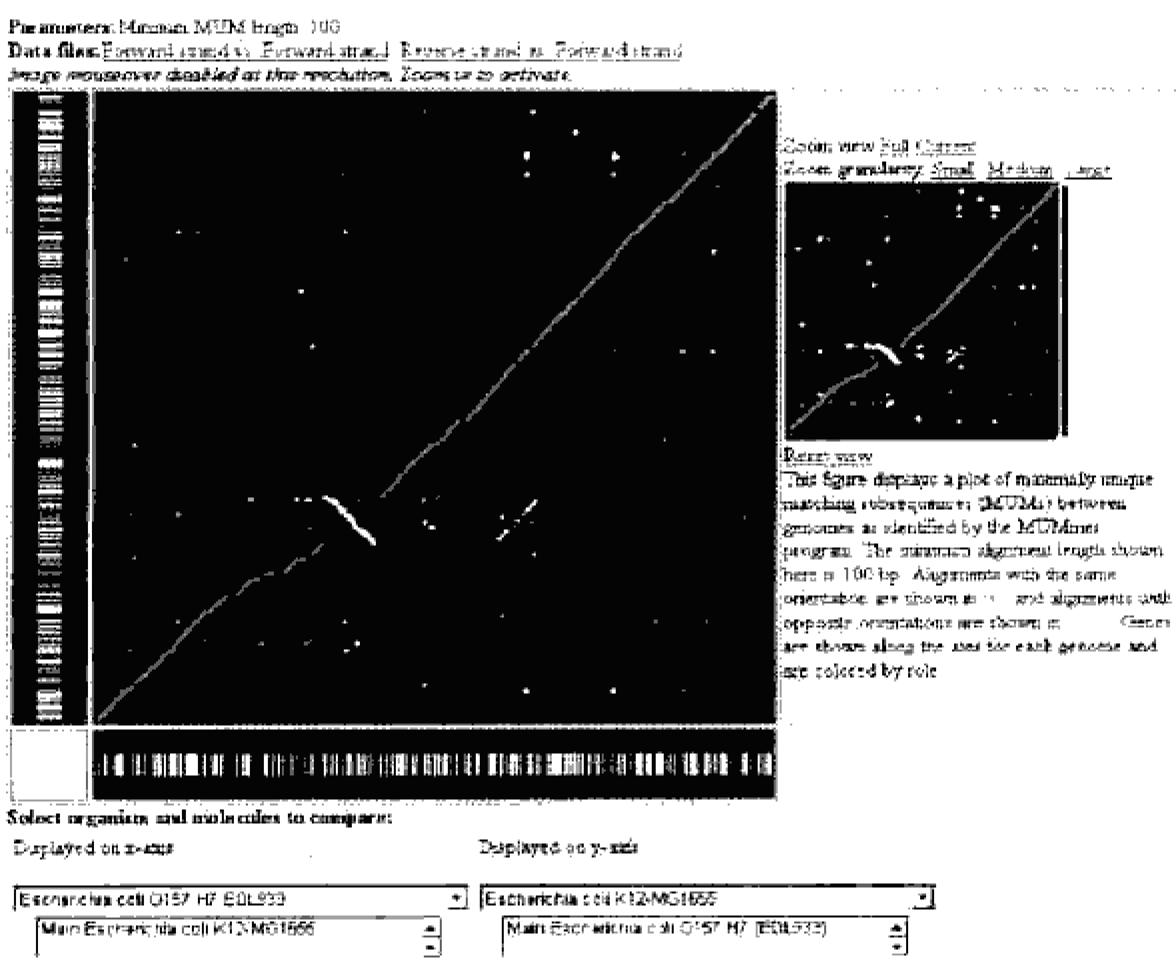


图 14-18 MUMmer 的输出结果是两个基因组之间的最大惟一匹配(MUMs)的点阵图。这个工具能够快速地描述两个基因组之间的联系，包括基因组 DNA 的相对方向和插入与删除的存在信息

株，会导致如出血性结肠炎之类的疾病)。这两个菌株在大约 45 亿年前发生分支（Reid 等，2000)。两个基因组都经过测序并比较（Blattner 等，1997；Perna 等，2001；Hayashi 等，2001；Eisen，2001)。大肠杆菌 O157：H7 大约比大肠杆菌 K12 长了 859000 个碱基对。这两个细菌有大约 4.1Mb 的共同基因组骨架，大肠杆菌 O157：H7 有另外的 1.4Mb 的序列（大部分通过水平基因转移得到)。MUMmer 的输出结果对找出两个基因组中的共同区域和

反向重复区域非常有用。

14.10 展望

实际上最近的一些对细菌和古细菌基因组的测序对微生物的各个方面都有非常重大的影响。我们总结全基因组测序的一些意义如下。

① 得到一个细菌和古细菌基因组的整个 DNA 序列后，我们对所有的基因和调控元件有了全面的理解。这就像得到了一个机器的零件列表，尽管我们还没有得到使用手册。

② 通过比较基因组学，我们能够学到“机器”组装和如何发挥作用的原理。

③ 我们能够通过比较基因组学来理解微生物种类的多样性。这样我们就能开始发现一些基因组组织的原理，并且我们也能比较一下致病菌和非致病菌。我们还能够看到来自同一种类的两个菌株之间基因组特点的巨大差异。

④ 我们开始深入了解基因和物种的进化过程。我们现在认为水平基因转移是一种微生物进化的驱动力，也能够研究基因复制和基因丢失。拥有全基因组对于明白什么基因组成一个生物和什么基因缺失了非常重要（Doolittle，1998）。

⑤ 全基因组序列给生物学研究提供了一个起点。

14.11 常见问题

随着细菌和古细菌基因组测序的完成，两个最重要的任务是找到基因和基因组注释。有几个原因导致鉴别基因比较困难：很难估计短开放阅读框是否真正被转录，原核生物（相对于真核生物）并不总是将 AUG 作为起始密码子。基因组注释是赋予预测蛋白质功能最关键的过程。通过计算得到的结果总被认为是假设的，需要实验验证。一般有以下几种错误（Brenner，1999；Mural，1999；Peri 等，2001）。

① 传递性错误：根据已知功能的同源基因错误地给一个基因赋予功能。

② 当短开放阅读框不转录的时候，将它误认为真正的基因。Devos 和 Valencia（2001）估计注释为普通功能的基因中大约 5%可能是错误的，但注释为特殊功能的基因中大约 33%可能是不正确的。

14.12 网络资源

TIGR 的综合微生物资源库（CMR）（http://www.tigr.org）为所有微生物基因组的研究提供了一个重要的起点。那个网站有一个专门的链接页面（http://www.tigr.org/CMR2/Links.shtml），提供了一个很好的链接到其他资源的通道。与微生物相关的链接是 http://www.microbes.info。这个站点有许多资源，包括对微生物学的一些介绍性文章。

14.13 问题讨论

1. 炭疽病菌株的致病性变化很大，你能够用哪种生物信息学方法来理解为什么会产生这些差异？

2. 你怎样估计一个细菌基因是否通过水平基因转移被整合到人类基因组中？某一个人类蛋白质和一组细菌蛋白质非常近源，在其他真核生物中却没有同源蛋白，对此现象有无其他解释？

3. 考虑大肠杆菌 K12 和 O157：H7 以及其他近源细菌对。它们经过了不同程度的水平基因转移，有不同的致病模式，并且两个菌株基因组大小有超过 100 万个碱基对的差异。物种的定义是什么？大肠杆菌是一个物种吗？

14.14 问题

1. 分析一个细菌基因组。任意选择一个已经被测序的细菌基因组。如果自己便于选择可选择大肠杆菌。先从 Entrez Genomes 开始，找到一个已知在真核生物中有同源物的基因。接着使用 TaxPlot，然后在 COG 站点上找一个已知在真核生物中有同源物的基因。除了 NCBI，还有两个非常好的关于已完成测序的基因组网站。

① 在 TIGR 网站上搜索这个细菌基因组。先登录 http://www.tigr.org，再（从左边的边框）点击 TIGR 数据库。在 TIGR 数据库中点击"已经完成的项目"，然后找到你需要的基因组。

② 在 Wellcome Trust Sanger Institute 网站上搜索这个细菌基因组：登录 http://www.sanger.ac.uk/Projects/。

2. 比较两个已经完成测序的基因组。从 Entrez Genomes 开始，先选择细菌，然后选择一个诸如 *Rickettsia prowazekii* 的一个物种。用 TaxPlot 进行一个三基因组比较。试着在图上点击其中一个点，然后只分析这个基因的功能组（"转录"）。在"未知功能"组中重复你的搜索过程，看看输出结果是否一样。

14.15 自测题

1. 一个典型的细菌基因组大概由多少个 DNA 碱基对组成？

(a) 20000bp；

(b) 200000bp；

(c) 2000000bp (2Mb)；

(d) 20000000bp (20Mb)

2. 一般来讲，为了研究细菌基因组需要特殊的生物信息学工具，这是因为

(a) 比起真核生物来，细菌用它们自己的基因密码；

(b) 比起真核生物来，细菌没有细胞核；

(c) 细菌很少有基因在真核生物中有同源物；

(d) 细菌 DNA 的基因含量和组织结构与真核生物有很大不同

3. 大肠杆菌基因组编码了大概 4300 个蛋白质编码基因。大肠杆菌内含子的总数量大概是

(a) 10 个；

(b) 430 个；

(c) 4300 个；

(d) 43000 个

4. 具有最小的原核生物基因组的生物可能是

(a) 嗜极生物；

(b) 病毒；

(c) 胞内细菌；

(d) 杆菌

5. 下面哪一个选项最好地描述了在一个

已经完全测序的古细菌蛋白质组中有多少个大肠杆菌蛋白质的同源物?

(a) PubMed;

(b) ColiBase;

(c) TaxPlot;

(d) MUMmer

6. 下面哪一个提供了最充分的证据证明一个大肠杆菌基因是通过水平基因转移被整合到大肠杆菌基因组中的?

(a) 比起其他大肠杆菌基因，这个基因的 GC 含量有很大差异;

(b) 比起其他大肠杆菌基因，这个基因使用密码子的频率有很大差异;

(c) 系统发生分析表明和大肠杆菌近源的 *proteobacteria* 没有这个基因;

(d) 上面中的任何一个

7. COG 数据库最主要的目的是

(a) 根据同源关系对已完成测序的原核生物基因组蛋白质进行分类;

(b) 为已完成测序的原核生物基因组提供多序列比对;

(c) 提供一个蛋白质功能分类系统;

(d) 根据原核生物中的保守家族来预测单个真核生物蛋白质的功能

14.16 推荐读物

关于细菌基因组学有大量的文献，其中 Casjens (1998) 对细菌基因组学作了一个非常棒的介绍，Joyce 等 (2002) 从基因组学角度发表了一个很有意义的关于细菌群落的综述。

参 考 文 献

Achenbach-Richter, L., Gupta, R., Stetter, K. O., and Woese, C. R. Were the original eubacteria thermophiles? *Syst. Appl. Microbiol.* **9**, 34-39 (1987).

Akman, L., Rio, R. V., Beard, C. B., and Aksoy, S. Genome size determination and coding capacity of *Sodalis glossinidius*, an enteric symbiont of tsetse flies, as revealed by hybridization to *Escherichia coli* gene arrays. *J. Bacteriol.* **183**, 4517-4525 (2001).

Akman, L., Yamashita, A., Watanabe, H., Oshima, K., Shiba, T., Hattori, M., and Aksoy, S. Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nat Genet.* **32**, 402-407 (2002).

Alm, R. A., et al. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**, 176-180 (1999).

Andersson, S. G., and Kurland, C. G. Reductive evolution of resident genomes. *Trends Microbiol.* **6**, 263-268 (1998).

Andersson, S. G., et al. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133-140 (1998).

Baytaluk, M. V., Gelfand, M. S., and Mironov, A. A. Exact mapping of prokaryotic gene starts. *Brief. Bioinform.* **3**, 181-194 (2002).

Bender, L., Ott, M., Marre, R., and Hacker, J. Genome analysis of *Legionella* ssp. by orthogonal field alternation gel electrophoresis (OFAGE). *FEMS Microbiol. Lett.* **60**, 253-257 (1990).

Bentley, S. D., et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141-147 (2002).

Blattner, F. R., et al. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453-1474 (1997).

Brenner, S. E. Errors in genome annotation. *Trends Genet.* **15**, 132-133 (1999).

Brochier, C., and Philippe, H. Phylogeny: A non-hyperthermophilic ancestor for bacteria. *Nature* **417**, 244 (2002).

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E., and Stanhope, M. J. Universal trees based on large combined protein sequence data sets. *Nat. Genet.* **28**, 281-285 (2001).

Bult, C. J., et al. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058-1073 (1996).

Casjens, S. The diverse and dynamic structure of bacterial genomes. *Annu. Rev. Genet.* **32**, 339-377 (1998).

Casjens, S., et al. A bacterial genome in flux: The twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Mol. Microbiol.* **35**, 490-516 (2000).

Chambaud, I., et al. The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acids Res.* **29**, 2145-2153 (2001).

Charon, N. W., and Goldstein, S. F. Genetics of motility and chemotaxis of a fascinating group of bacteria: The Spirochetes. *Annu. Rev. Genet.* **36**, 47-73 (2002).

Cole, S. T., et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537-544 (1998).

Cole, S. T., et al. Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007-1011 (2001).

Dandekar, T., et al. Re-annotating the *Mycoplasma pneumoniae*

genome sequence: Adding value, function and reading frames. *Nucleic Acids Res.* **28**, 3278–3288 (2000).

Deckert, G., et al. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus. Nature* **392**, 353–358 (1998).

Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999a).

Delcher, A. L., et al. Alignment of whole genomes. *Nucleic Acids Res.* **27**, 2369–2376 (1999b).

Delcher, A. L., Phillippy, A., Carlton, J., and Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478–2483 (2002).

DeLong, E. F. Everything in moderation: Archaea as "non-extremophiles." *Curr. Opin. Genet. Dev.* **8**, 649–654 (1998).

DeLong, E. F., and Pace, N. R. Environmental diversity of bacteria and archaea. *Syst. Biol.* **50**, 470–478 (2001).

de Villiers, E. P., Brayton, K. A., Zweygarth, E., and Allsopp, B. A. Genome size and genetic map of *Cowdria ruminantium. Microbiology* **146**, 2627–2634 (2000).

Devos, D., and Valencia, A. Intrinsic errors in genome annotation. *Trends Genet.* **17**, 429–431 (2001).

Dobell, C. *Antony van Leeuwenhoek and his "little animals".* Harcourt, Brace and Company, New York, 1932.

Doolittle, R. F. Microbial genomes opened up. *Nature* **392**, 339–342 (1998).

Eisen, J. A. Horizontal gene transfer among microbial genomes: New insights from complete genome analysis. *Curr. Opin. Genet. Dev.* **10**, 606–611 (2000).

Eisen, J. A. Gastrogenomics. *Nature* **409**, 463, 465–466 (2001).

Eisenstein, B. I., and Schaechter, M. Normal microbial flora. In M. Schaechter, N. C. Engleberg, B. I. Eisenstein, and G. Medoff (Eds.), *Mechanisms of Microbial Disease.* Lippincott Williams and Wilkins, Baltimore, MD, 1999, Chapter 20.

Ermolaeva, M. D., White, O., and Salzberg, S. L. Prediction of operons in microbial genomes. *Nucleic Acids Res.* **29**, 1216–1221 (2001).

Feng, D. F., Cho, G., and Doolittle, R. F. Determining divergence times with a protein clock: Update and reevaluation. *Proc. Natl. Acad. Sci. USA* **94**, 13028–13033 (1997).

Fleischmann, R. D., et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).

Fox, G. E., et al. The phylogeny of prokaryotes. *Science* **209**, 457–463 (1980).

Fraser, C. M., et al. The minimal gene complement of *Mycoplasma genitalium. Science* **270**, 397–403 (1995).

Fraser, C. M., et al. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi. Nature* **390**, 580–586 (1997).

Fraser, C. M., et al. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**, 375–388 (1998).

Fraser, C. M., Eisen, J. A., and Salzberg, S. L. Microbial genome sequencing. *Nature* **406**, 799–803 (2000).

Glaser, P., et al. Comparative genomics of *Listeria* species. *Science* **294**, 849–852 (2001).

Glass, J. I., et al. The complete sequence of the mucosal pathogen *Ureaplasma urealyticum. Nature* **407**, 757–762 (2000).

Graur, D., and Li, W.-H. *Fundamentals of Molecular Evolution.* Sinauer Associates, Sunderland, MA, 2000, p. 481.

Gupta, R. S., and Griffiths, E. Critical issues in bacterial phylogeny. *Theor. Popul. Biol.* **61**, 423–434 (2002).

Hayashi, T., et al. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* **8**, 11–22 (2001).

Heidelberg, J. F., et al. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae. Nature* **406**, 477–483 (2000).

Himmelreich, R., et al. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae. Nucleic Acids Res.* **24**, 4420–4449 (1996).

Huber, H., et al. A new phylum of *Archaea* represented by a nanosized hyperthermophilic symbiont. *Nature* **417**, 63–67 (2002).

Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Genome Biol.* 3(2) (2002).

Hugenholtz, P., Goebel, B. M., and Pace, N. R. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* **180**, 4765–4774 (1998).

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

Itaya, M. An estimation of minimal genome size required for life. *FEBS Lett.* **362**, 257–260 (1995).

Joyce, E. A., Chan, K., Salama, N. R., and Falkow, S. Redefining bacterial populations: A post-genomic reformation. *Nat. Rev. Genet.* **3**, 462–473 (2002).

Kalman, S., et al. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis. Nat. Genet.* **21**, 385–389 (1999).

Kawarabayasi, Y., et al. Complete genome sequence of an aerobic hyperthermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res.* **6**, 83–101, 145–152 (1999).

Klenk, H. P., et al. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus. Nature* **390**, 364–370 (1997).

Koonin, E. V. Genome sequences: Genome sequence of a model prokaryote. *Curr. Biol.* **7**, R656–R659 (1997).

Koonin, E. V., Makarova, K. S., and Aravind, L. Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu. Rev. Microbiol.* **55**, 709–742 (2001).

Kunst, F., et al. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis. Nature* **390**, 249–256 (1997).

Ma, J., Campbell, A., and Karlin, S. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J. Bacteriol.* **184**, 5733–5745 (2002).

Malloff, C. A., Fernandez, R. C., Dullaghan, E. M., Stokes, R. W., and Lam, W. L. Two-dimensional display and whole genome comparison of bacterial pathogen genomes of high G+C DNA content. *Gene* **293**, 205–211 (2002).

May, B. J., et al. Complete genomic sequence of *Pasteurella multocida*, Pm70. *Proc. Natl. Acad. Sci. USA* **98**, 3460–3465 (2001).

Miller, W. Comparison of genomic DNA sequences: Solved and unsolved problems. *Bioinformatics* **17**, 391–397 (2001).

Mural, R. J. Current status of computational gene finding: a perspective. *Methods Enzymol.* **303**, 77–83 (1999).

Mural, R. J. ARTEMIS: A tool for displaying and annotating DNA sequence. *Brief. Bioinform.* **1**, 199–200 (2000).

Mushegian, A. R., and Koonin, E. V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. USA* **93**, 10268–10273 (1996).

Nelson, K. E., et al. Evidence for lateral gene transfer between *Archaea* and bacteria from genome sequence of *Thermotoga maritima. Nature* **399,** 323–329 (1999).

Nolling, J., et al. Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum. J. Bacteriol.* **183,** 4823–4838 (2001).

Ogata, H., et al. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii. Science* **293,** 2093–2098 (2001).

Owen, Richard. Lectures on the Comparative Anatomy and Physiology of the Invertebrate Animals. Longman, Brown, Green, and Longmans, London, 1843.

Parkhill, J., et al. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar *typhi* CT18. *Nature* **413,** 848–852 (2001a).

Parkhill, J., et al. Genome sequence of *Yersinia pestis,* the causative agent of plague. *Nature* **413,** 523–527 (2001b).

Peri, S., Ibarrola, N., Blagoev, B., Mann, M., and Pandey, A. Common pitfalls in bioinformatics-based analyses: Look before you leap. *Trends Genet.* **17,** 541–545 (2001).

Perna, N. T., et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409,** 529–533 (2001).

Perrière, G., Duret, L., and Gouy, M. HOBACGEN: Database system for comparative genomics in bacteria. *Genome Res.* **10,** 379–385 (2000).

Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K., and White, O. The Comprehensive Microbial Resource. *Nucleic Acids Res.* **29,** 123–125 (2001).

Ponting, C. P. Plagiarized bacterial genes in the human book of life. *Trends Genet.* **17,** 235–237 (2001).

Read, T. D., et al. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.* **28,** 1397–1406 (2000).

Read, T. D., et al. Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis. Science* **296,** 2028–2033 (2002).

Reid, S. D., Herbelin, C. J., Bumbaugh, A. C., Selander, R. K., and Whittam, T. S. Parallel evolution of virulence in pathogenic *Escherichia coli. Nature* **406,** 64–67 (2000).

Reysenbach, A. L., and Shock, E. Merging genomes with geochemistry in hydrothermal ecosystems. *Science* **296,** 1077–1082 (2002).

Salzberg, S. L., White, O., Peterson, J., and Eisen, J. A. Microbial genes in the human genome: Lateral transfer or gene loss? *Science* **292,** 1903–1906 (2001).

Schaechter, M. Introduction to the pathogenic bacteria. In M., Schaechter, N. C., Engleberg, B. I., Eisenstein, and G., Medoff (Eds.), *Mechanisms of Microbial Disease.* Lippincott Williams and Wilkins, Baltimore, MD, 1999, Chapter 10.

Serres, M. H., et al. A functional update of the *Escherichia coli* K-12 genome. *Genome Biol.* **2(9)**:research 0035.1–0035.7 (2001).

Sherratt, D. Divide and rule: The bacterial chromosome. *Trends Genet.* **17,** 312–313 (2001).

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407,** 81–86 (2000).

Shirai, M., et al. Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. *Nucleic Acids Res.* **28,** 2311–2314 (2000).

Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D., and Krogh, A. On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.* **17,** 425–428 (2001).

Stanhope, M. J., et al. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* **411,** 940–944 (2001).

Stephens, R. S., et al. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis. Science* **282,** 754–759 (1998).

Stover, C. K., et al. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406,** 959–964 (2000).

Sun, L. V., et al. Determination of *Wolbachia* genome size by pulsed-field gel electrophoresis. *J. Bacteriol.* **183,** 2219–2225 (2001).

Tamas, I., et al. 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296,** 2376–2379 (2002).

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–637 (1997).

Tatusov, R. L., et al. The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29,** 22–28 (2001).

Tettelin, H., et al. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287,** 1809–1815 (2000).

Tomb, J. F., et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori. Nature* **388,** 539–547 (1997).

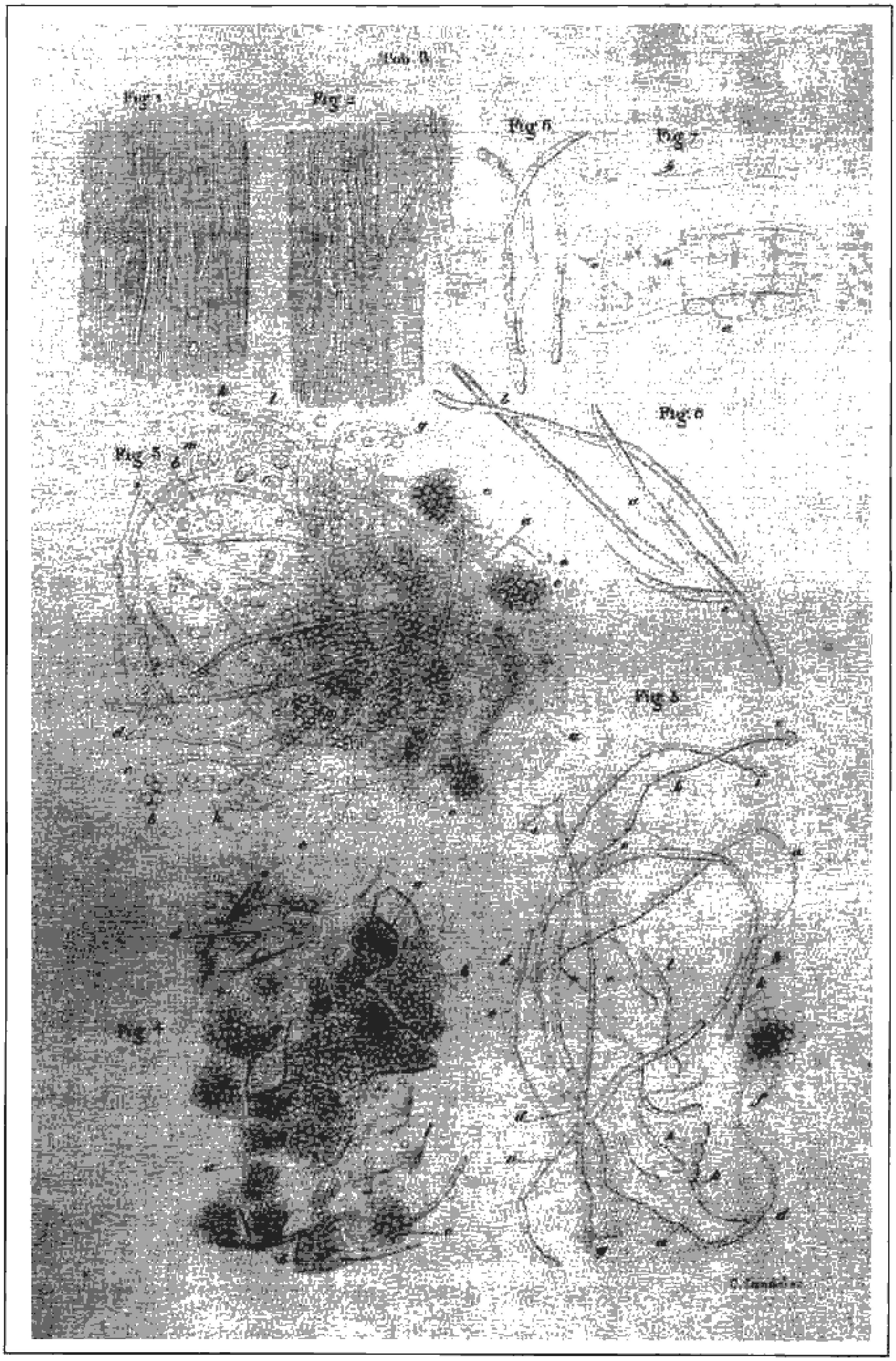
White, O., et al. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286,** 1571–1577 (1999).

Woese, C. R., and Fox, G. E. The concept of cellular evolution. *J. Mol. Evol.* **10,** 1–6 (1977).

Woese, C. R. On the evolution of cells. *Proc. Natl. Acad. Sci. USA* **99,** 8742–8747 (2002).

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L., and Koonin, E. V. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* **1,** 8 (2001).

Wood, V., et al. The genome sequence of *Schizosaccharomyces pombe. Nature* **415,** 871–880 (2002).



已知有200多个真菌类是对人类致病的，威胁着数百万人。从希腊、古罗马时代到19世纪中期人们只知道两种真菌病：癣菌病（癣）和鹅口疮（口腔念珠菌病）（Ainsworth，1993）。癣菌病是由小孢子癣菌、毛癣菌和表皮癣菌这几类真菌导致的。念珠菌病则是由于白色假丝酵母和其他假丝酵母类而引起的。这张图（摘自Kuchenmeister，1857）展示的是念珠菌，这在那时候被称作白色粉孢。

第 15 章

真核基因组：真菌基因组

15.1 引言

真菌学（来自希腊单词 mukes，“真菌”）就是对真菌的研究。霉菌病是一种由真菌导致的疾病。后缀-mycota 是指真菌；真菌界也被称为真的真菌界（Eumycota）（“真的真菌”）。

形态学上，真菌的特点就是生长有分支的菌丝（细丝）。在加州大学伯克利分校的古生物学博物馆中，有一个关于真菌的介绍，包括许多物种的照片（http://www.ucmp.berkeley.edu/fungi/fungi.html）。美国国家历史博物馆（纽约）也有一个关于真菌的概述（http://ology.amnh.org/biodiversity/treeoflife/pages/fungi.html）。

根据 Whittaker 的分类系统，生命有五界：无核原生生物（原核）、原生生物、动物、真菌和植物。我们已经在第 14 章探讨了原核生物。在本章我们开始通过学习真菌来探索真核生物。这群多样化、有趣的生物最初在 15 亿年前与植物和动物共有同一个祖先（Wang 等，1999，在第 16 章讨论）。我们可以想像真菌就像蘑菇一样已经被生物学家研究过了。令人惊奇的是，真菌比起植物来与动物更为亲近。在第 16 章我们将会接着学习整个真核生物界，包括动物、植物和各种原生动物。然后我们会讨论人类（第 17 章）。

我们从真菌开始对真核生物的学习。第一个被完全测序的真核基因组是一个具有 1300 万碱基对的真菌——出芽酵母的基因组，比起人类基因组（30 亿碱基对，或千兆碱基对）显得非常小，并且它的大小也只有一个标准细菌基因组的几倍大。这种酵母已经作为一种真核模式生物来用于基因研究，因为它增长速度快，很容易进行基因修饰，并且它的许多细胞功能与多细胞生物和其他真核生物一样是保守的。最近，它更成为了基因组学和生物信息学研究的一种模式生物。它大概有 6000 多个基因，每个基因都被描绘、删除、过表达并用不同的实验来描述其功能（见下文）。

真菌学（来自希腊语 mukes，“真菌”）就是真菌的学科。霉菌病是一种由真菌导致的疾病。后缀-mycota 表示真菌；真菌界也被叫做 Eumycota 界。

本章首先对真菌进行概述。然后我们将会介绍出芽酵母的生物信息学方法。最后，我们会介绍其他真菌基因组的测序和真菌比较基因组学的早期课程。

真菌的分类标准和类型。真菌是丝状的（如霉菌）或者单细胞

（如酵母中的出芽酵母）的真核生物。主要的细菌分类标准在于形态（如超微结构）、生物化学性质（如生长性质或者细胞壁组成）和分子序列数据（DNA、RNA 和蛋白质序列）。大多数真菌是需氧的，并且所有的都是异养生物，需要摄入食物。真菌经常是非常坚硬的，形成由壳聚糖组成的孢子，在整个生命过程中都无法移动。它们在生态系统中为降解生物废物起了重要的作用。真菌是人类、其他动物和植物最重要的致病媒介。真菌在发酵工业中也起着重要作用：霉菌 *Rhizopus nigricans* 用于如可的松之类的类固醇的大规模生产，*Penicillium chrysogenum* 能生产抗生素——青霉素。

在生命树中各个物种之间的关系已经用基于小亚单位核糖 RNA 的系统发生分析来进行描述（图 12-1）。在另外一种与之相反的方法中，W. F. Doolittle 和他的同事根据链接氨基酸序列定义了一个真核生物的系统发生关系，这些氨基酸序列来自下面几个蛋白质：延伸因子-1α、肌动蛋白、α-微管蛋白和 β-微管蛋白（Baldaufet 等，2000）。从发生树上那部分可以看出真菌形成了和多细胞动物并列的一个单元进化枝（图 15-1）。考虑到许多真菌都具有明显的简单单细胞动物性质，真菌和动物之间这种近源关系在一定程度上让人感到惊奇。然而，真菌和动物有很多地方具有相似性。壳多糖是真菌细胞壁主要的成分，它也是节肢动物外骨骼的组成成分（植物细胞壁采用纤维素）。酵母的许多基本生物过程（如细胞周期控制、DNA 修复和胞内囊泡转运）都和哺乳动物细胞一样保守。

我们将会在第 16 章详细地来讨论这个进化树（图 16-1）
参见框 15-1 关于真菌分类学的讨论。

在阿波罗登月计划之前，有些宇航员被真菌感染了，这些真菌包括 *Aspergillus fumigatus*、*Candida albicans* 和 *Trichophyton rubrum*（http://lsda.jsc.nasa.gov/scripts/cf/exper.cfm?exp_index=372）。在指挥舱的密闭环境下，有些致病菌缺少了正常微生物的竞争，在登月过程中大量繁殖。找到和治疗真菌感染在整个登月计划中都受到了关注。

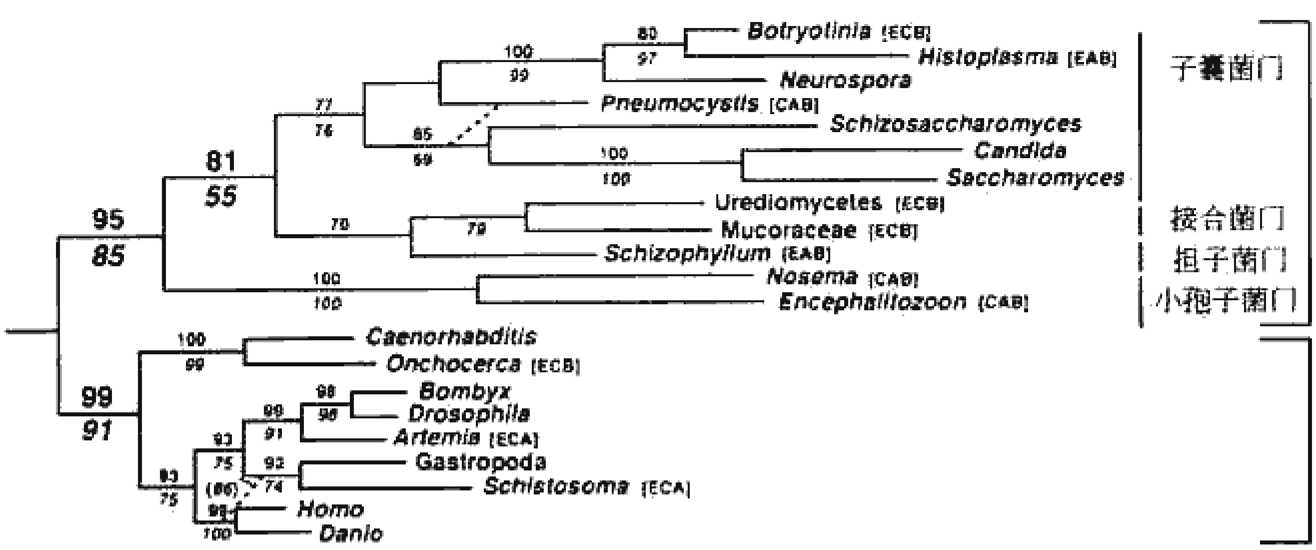


图 15-1 真菌的系统发生分析表明它们是和多细胞生物（动物）属于姐妹分支。由 Baldauf 等得到的这个发生树是对真核生物分析的详细表示（2000）。这个发生树是通过 4 个蛋白质序列的多序列比对得到的，这 4 个蛋白质为延长因子 1α（EF-1α）（在发生树中简写为 E）、肌动蛋白（C）、α 微管蛋白（A）和 β 微管蛋白。小孢子菌原来被划归为底部分支的真核生物，现在却被归为真菌这一组

框 15-1 真菌分类

在1995年大概有70000个真菌种类被发现，尽管总的真菌种类数目估计至少有150万。这些真菌分成了四个门：子囊菌、担子菌、壶菌和接合菌（Fuarro等，1999）。子囊菌包括酵母、蓝绿霉、块菌和地衣；已知的大概有30000种，包括曲霉属、念珠菌属、隐菌属（不确定）、组织胞浆菌属、链孢霉属、酵母属。担子菌包括锈菌、黑穗病菌和蘑菇；它们靠其称为孢子台的棒状复制结构与其他真菌属相区别。壶菌，有时也被划分为原始有核界（Margulis和Schwartz，1998），包括别霉菌和*polyphagus*。最后，接合菌属的真菌都没有隔膜（隔板），一般以腐烂的蔬菜为食，包括球菌属、白霉属和酒曲菌属。

子囊菌具有比较特殊的意义，因为酵母属于这一门。它又可以分为四类：半子囊菌（如啤酒酵母）、真囊菌（如*Neurospora crassa*）、Loculoascomycetae（如*Elsinoe proteae*）和Laboulbeniomycetae（昆虫的寄生虫）。

真菌除了上面的四个门，有些也在Chromista界（Hyphochytriomycota、labyrinthulomycota和Oomycota）和原生动物界（Acrasiomycota、Dictyosteliomycota、Mxomycota和Plasmodiophoromycota）被发现。真菌在不同的界中都被发现表明以下观点：在分类学历史上真菌就是按照形态学标准来分的，而形态学标准有时候并不与分子标准相一致。

15.2 出芽酵母介绍

15.2.1 啤酒酵母

从Anton van Leeuwenhoek时代（1632～1723）开始，酵母就被认为是没有生命的化学物质。Theodor Schwann（1810～1882）和Baron Charles Cagniard-Latour（1777～1859）在1836～1837年独立发现了酵母是由活细胞组成的。Schwann研究了发酵酵母并称它们为*Zuckerpilz*（糖真菌），从而产生了*Saccharomyces*这个词（Bulloch，1938）。

*Saccharomyces cerevisiae*经常用来称呼“出芽酵母”以便和“裂殖酵母”（*Schizosaccharomyces pombe*）区别。裂殖酵母是第二个测序的真菌基因组（见下文）。出芽酵母是一种单细胞生物，在复制的过程中出芽分裂。

啤酒酵母的序列是在1996年4月24日首次发布的。

啤酒酵母是第一个至少在1万年前就能被人工培育的真菌，它一般被称为酿造酵母或面包酵母，能把葡萄糖发酵成为乙醇和二氧化碳。100多年来，研究者们从生化、基因、分子和细胞生物学角度来对它进行研究。因为它的许多性质在人类细胞中也同样保守，酵母成为了一个对于基础研究非常重要的生物。

15.2.2 酵母基因组测序

目前，相对较小的基因组用全基因组“乌枪”法（参见第12章）测序。相比较而言，酵母基因组是在20世纪90年代中期通过染色体来测序的，由来自全世界超过600名研究者共同完成的（Mewes等，1997）。这个工作分成好几个阶段进行：第一，它的16个染色体的粗略物理图谱用限制性切割酶构建出来。第二，一个大约10kb的基因组DNA插入文库在λ噬菌体中构建起来，这些插入DNA用限制酶印迹识别出来。然后将重叠的插入DNA通过计算机分析确定克隆DNA，整合成为16个大的DNA序列片段。然后从中选出最少重复而且能够覆盖整个基因组的一套重复片段，分配到各个实验室用不同的方法进行测序。每个实验室都要保证其准确率。最后，将这些单个的序列收集起来集中编辑，整合成为16个全染色体，然后用标准命名法来进行编注［最后的错误率在0.003%以下（Mewes等，1997）］。今天，这个方法已经被认为费力、低效并且昂贵。然而，这个合作却是非常成功的。

15.2.3 出芽酵母基因组的特征

啤酒酵母基因组由 16 个染色体组成共有约 13Mb 的 DNA。有了基因组全序列，啤酒酵母的物理图谱（直接由 DNA 测序决定）就和其基因图谱（通过四分体分析来推断出基因之间的距离）统一起来了（Cherry 等，1997）。最终的全序列由 300000 个独立片段整合而成（Mewes 等，1997）。啤酒酵母的一些序列特征在表 15-1 中列出来了，这些特征是根据最初的基因组注释得到的（Goffeau 等，1996）。

表 15-1 最初注释的啤酒酵母基因组特征

特 征	数 量	特 征	数 量
序列长度	12068kb	开放阅读框中的内含子	220
重复片段长度	1321kb	*UTRs* 中的内含子	15
总长度	13389kb	完整的 *Ty* 元件	52
开放阅读框	6275	tRNA 基因	275
有问题的蛋白质	390	snoRNA 基因	40
假想蛋白	5885		

注：1. 根据 Goffeau 修改得到（1996）这个表格的升级版本能够在 MIPS 综合酵母基因组数据中得到（http://mips.gsf.de/proj/yeast/CYGD/db/index.html）。

2. snoRNA 表示小核仁 RNA，tRNA 表示转运 RNA，*Ty* 表示逆转座子，*UTRs* 表示未翻译区。

酵母基因组的一个显著特点就是它的高基因密度（大约每 2000 个碱基一个基因）。然而细菌的基因密度大约为 1000 个碱基中一个基因，大多数高级真核生物的基因密度比较低。相比较而言，在裂殖酵母中，40%的基因有内含子（见下文）。没有内含子的特性使得啤酒酵母成为一个有吸引力的模式生物，用来从基因组 DNA 中鉴别出基因。表 15-2 和表 15-3 中所列出的是啤酒酵母中最为常见的蛋白质家族和蛋白质结构域。EBI 提供了这种生物和其他生物（图 15-2）一系列不同的蛋白质组学分析，如蛋白质长度分析（图 15-3）。

表 15-2 啤酒酵母的 15 个最常见蛋白质家族

InterPro ID	匹配的蛋白质数目	名 称
IPR000719	115	蛋白激酶
IPR005828	53	一般物质转运蛋白
IPR001042	52	TYA 转座子蛋白
IPR001969	40	真核/病毒天冬氨酸蛋白酶，活化位点
IPR000379	37	脂酶/脂肪酶/硫脂酶，活化位点
IPR001806	34	Ras GTP 酶超家族
IPR001993	34	线粒体物质运输蛋白
IPR005834	28	类脱卤酶的水解酶
IPR000992	27	压力引起的 SRP1/TIP1 蛋白家族
IPR002293	25	氨基酸/多氨基转运蛋白，家族Ⅰ
IPR001142	23	酵母膜蛋白 DUP
IPR002085	21	含锌乙醇脱氢酶超家族
IPR001394	18	泛素硫脂酶，家族Ⅱ
IPR001757	17	ATP 酶，E1-E2 型
IPR001163	16	小核蛋白(Sm 蛋白)

注：来源于 http://www.ebi.ac.uk/proteome/（2002 年 11 月）。

表 15-3 已经完成测序的啤酒酵母基因组的 15 个最常见功能域

InterPro ID	每个基因组的匹配数	蛋白质数目	名 称
IPR001650	144	71	解旋酶 C 端
IPR003593	70	58	AAA ATP 酶
IPR001138	238	57	真菌转录调控蛋白，N 端
IPR000504	301	55	结合区域 RNP-1(RNA 识别模体)
IPR000822	378	53	锌指，C2H2 型
IPR000822	43	42	小 GTP 结合蛋白域
IPR001841	165	40	锌指，RING
IPR001584	37	37	整合酶，催化域
IPR001849	77	29	pleckstrin homology(PH)域
IPR001452	144	25	SH3 域
IPR004841	22	22	透性酶中的域
IPR001623	79	21	热激蛋白 DnaJ，N 端
IPR004843	41	21	金属磷脂酶
IPR000051	20	20	SAM(和一些其他核苷酸)结合模体
IPR000629	20	20	ATP 依赖的解旋酶，DEAD 盒

注：来自欧洲生物信息学研究所（EBI）蛋白质组学分析网站（http://www.ebi.ac.uk/proteome/，2002 年 11 月）。

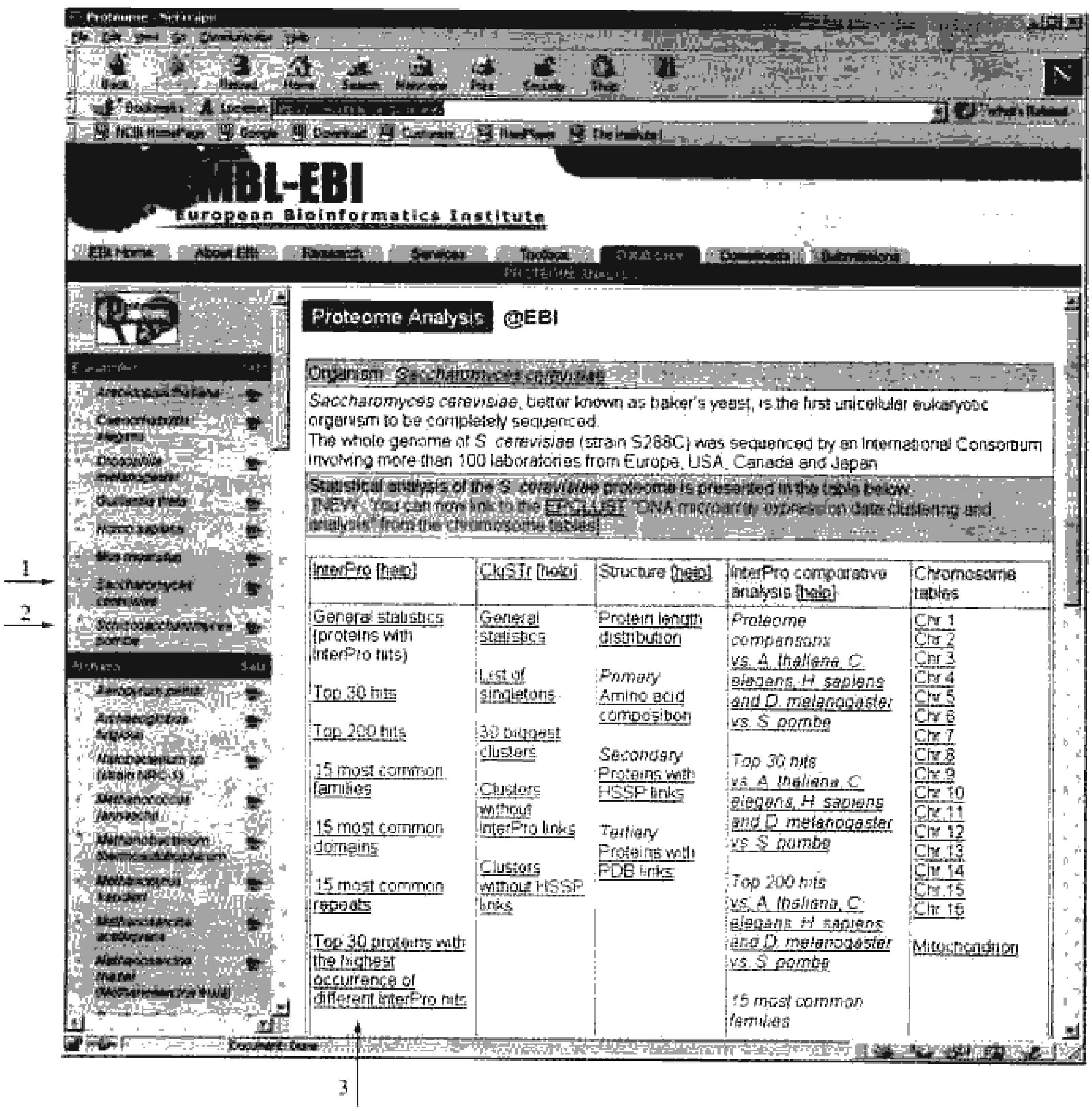


图 15-2 欧洲生物信息学研究所提供了针对啤酒酵母和许多其他生物的蛋白质组分析工具（http://www.ebi.ac.uk/proteome/）。在左边框中选择一种生物（箭头 1 和箭头 2 分别表示两个真菌蛋白质组）。还有许多蛋白质组可以选择，包括 InterPro 等（箭头 3）

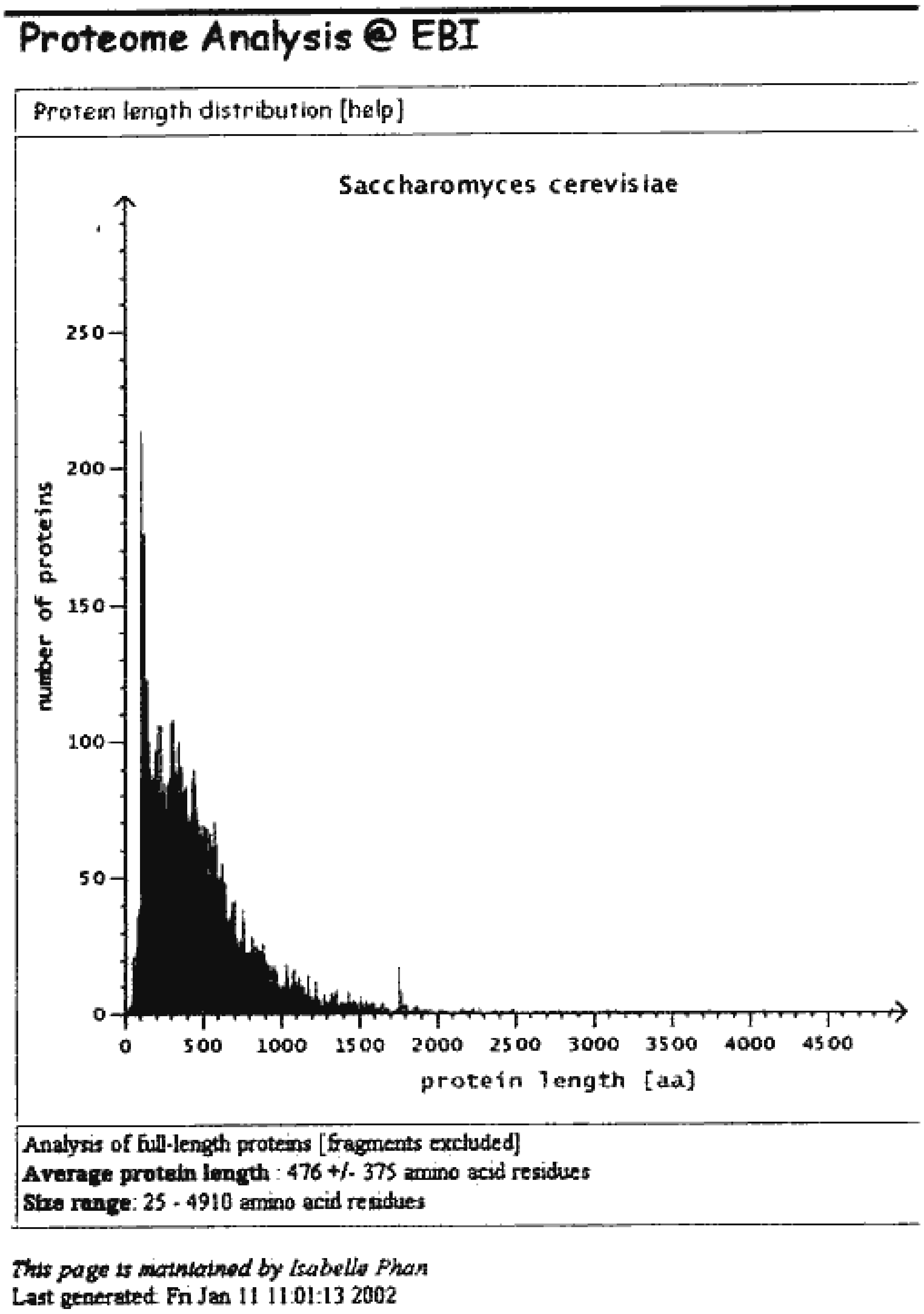


图 15-3 EBI 蛋白质组分析中包含了啤酒酵母中的蛋白质数量和蛋白质序列长度关系图。在预测的是小蛋白质情况下（如小于 100 个密码子编码的蛋白），一定要确认一下这个蛋白质在体内是否进行转录和翻译。否则编码这个蛋白的很可能是一个偶然发生的开放阅读框，并没有生物学意义

在最初注释基因组序列时，一共有 6275 个预测的开放阅读框（ORFs）。开放阅读框的要求为长度大于 100 个密码子（300 个核苷酸），这样一个对应的蛋白质至少有 11500u。其中，390 个开放阅读框（表 15-1 中列出）被认为是有问题的，因为它们的长度比较短而且不大可能去编码蛋白质（Dujon 等，1994）。根据“密码子适合指数”<0.01，有问题的开放阅读框表现为不被密码子使用所偏好。

> 根据定义，所有的开放阅读框以起始密码子（典型的是 AUG，编码甲硫氨酸）开始，以终止密码子（一般为 UAG、UAA 或 UGA）结束。

啤酒酵母中有多少蛋白质编码基因呢？是否有可能短的开放阅读框也编码真正的蛋白质呢？这些问题对于我们理解每一个真核生物基因组都是基本的。在注释酵母基因组的时候，可能会有假阳性（识别出不能真正编码基因的开放阅读框）和假阴性（真正的短开放阅读框基因没有被注释出来）。共有 40000 个开放阅读框比 20 个密码子长（Mackiewicz 等，2002）。随意设置一个 100 个密码子的域值，低于这个域值时，有很多开放阅读框满足密码子适合指数 >0.11 并且不和较长的开放阅读框重叠的标准（Harrison 等，2002）。决定一个开放阅读框是否为编码基因最主要的标准是，①是否在其他物种中保守并且/或者②存在基因表达的实验证据（见第 12 章）。对于啤酒酵母，Winzeler 和他的同事同

时运用寡核苷酸的基因表达谱和质谱证实了 138 个原来没有被注释的转录基因和 50 个原来没有被注释的翻译基因（Oshiro 等，2002）。Michael Snyder 和他的同事们结合了基因表达谱、转座子调节的基因捕捉法（见下文）和同源查找三个方法鉴别出了 137 个基因（Kumar 等，2002a）。

除了蛋白质编码基因，还有很多转录基因，这些基因编码了功能 RNA 分子却不接着翻译成为蛋白质。除了在表 15-1 中列出的 275 个 tRNA 基因，还有 140 个串联的重复 rRNA 基因拷贝，以及小片段 PNA（snoRNA）（Lowe 和 Eddy，1999）和其他种类的 RNA。

人类基因组中超过一半的部分是由可转座元件构成的，我们将在第 16 章和第 17 章中详细加以讨论。

MIPS 综合性酵母基因组数据库是啤酒酵母基因组信息的储存中心（Mewes 等，2002）。在这个数据库中根据 FASTA 分值用好几种方法来描述开放阅读框（表 15-4）。相似性较差的蛋白质的 FASTA 分值在 100～200。高度相似的蛋白质的 FASTA 分值在 200 和 1/3 的“自我打分值”（蛋白质和自己比对的分值）之间。有问题的开放阅读框一般有较低的密码子适合指数，同较长的或者已知的开放阅读框并且/或者与其他开放阅读框没有相似性。

表 15-4 MIPS 综合酵母基因组数据库中的酵母开放阅读框

分类	蛋白质数目	分类	蛋白质数目
已知蛋白	3400	没有相似性的	516
与已知蛋白有强相似性的	230	有问题的开放阅读框	472
与已知蛋白相似或具有低相似性的	825	总计	6450
与已知蛋白相似的	1007		

注：1. 这个数据库列出了所有的开放阅读框种类。
2. 来源于 http://mips.gsf.de/proj/yeast/CYGD/db/index.html（2002 年 11 月）。

啤酒酵母基因组编码了 52 个完整的逆转座子（共有 *Ty1*、*Ty2*、*Ty3*、*Ty4* 和 *Ty5* 五种）。有一些内源的元件（如逆转录病毒的元件）调控转座（例如，插入到一个新的基因组位置）（Roth，2000）。它们被长末端重复片段（*LTRs*）侧面攻击，这些长末端重复片段将逆转座子整合到新的基因组位点上去。逆转座子形成了所有真核基因组的基因分布。

十二号染色体（NC_001144）有 1078173 个碱基对。为了在 MIPS 上面浏览十二号染色体图谱，可以登录 http://mips.gsf.de/proj/yeast/CYGD/db/index.html。对于 SGD，可以访问 http://genome-www.stanford.edu/Saccharomyces/MAP/GENOMICVIEW/Genomic View.html。我们将在后文详细讨论这些站点。

15.2.4 探索典型的酵母染色体

你可以通过一些网站得到任何一个啤酒酵母染色体的 DNA 序列。我们用 NCBI 的基因组分类：点击左边条带上面的“最主要的生物”然后选择这 16 个染色体中的任意一个（图 15-4）。十二号染色体刚刚超过 1Mb（Johnston 等，1997）（图 15-5），我们来考察一下这个染色体的特殊性质。像原核生物（第 14 章）一样，NCBI 也提供了类似 TaxPlot、COG 等的相关特征。你也能够在一些专门的酵母数据库中查看酵母的十二号染色体，如 MIPS 综合酵母基因组数据库（图 15-6）和酵母基因组数据库（SGD）（Dwight 等，2002）

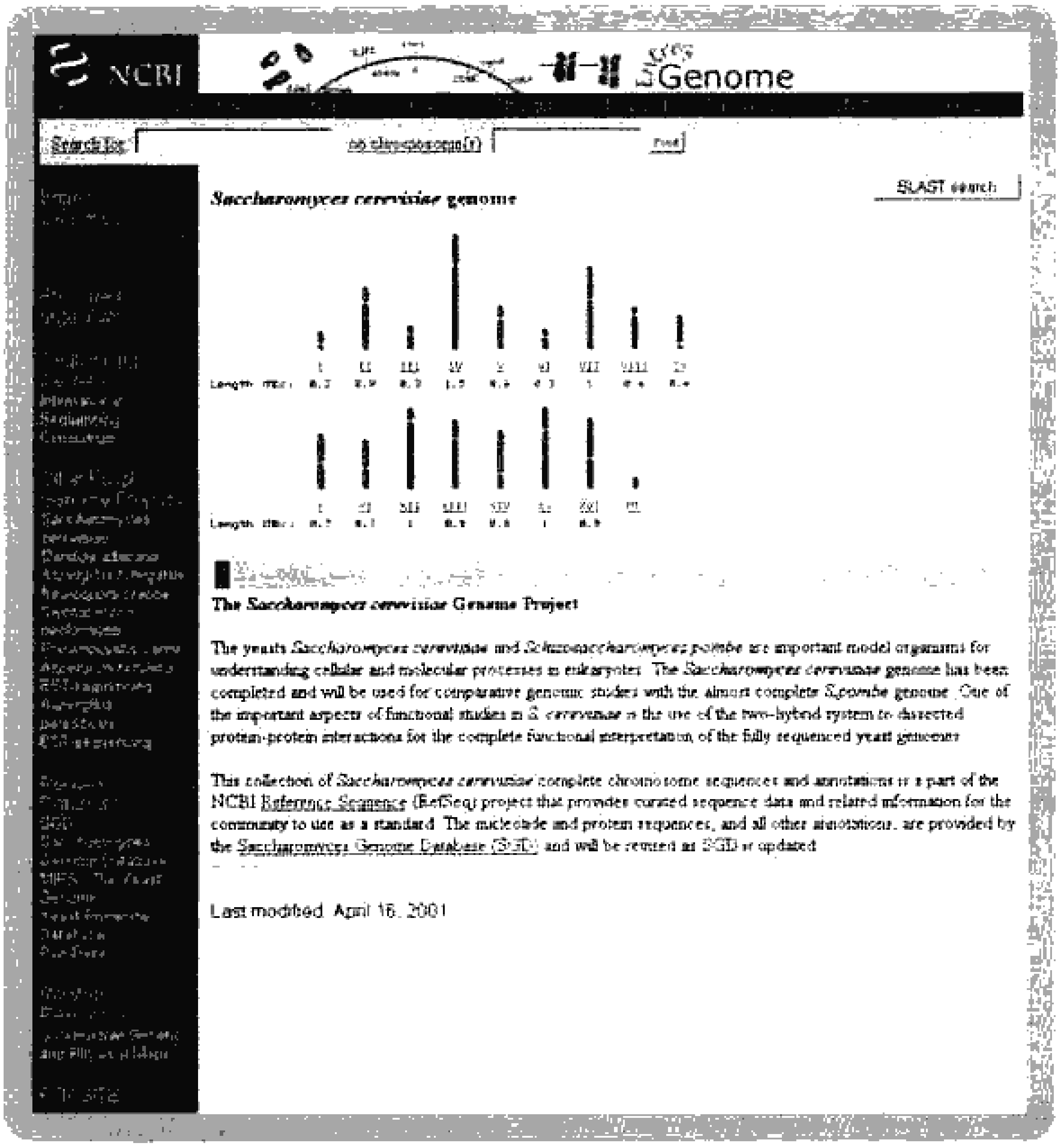


图 15-4 NCBI 上的基因组查询包含这个关于啤酒酵母的页面，对 16 个染色体中的每一个都可以进行独立研究。左边的框中（黑底）是一些关于啤酒酵母和其他真菌的主要网上资源的链接

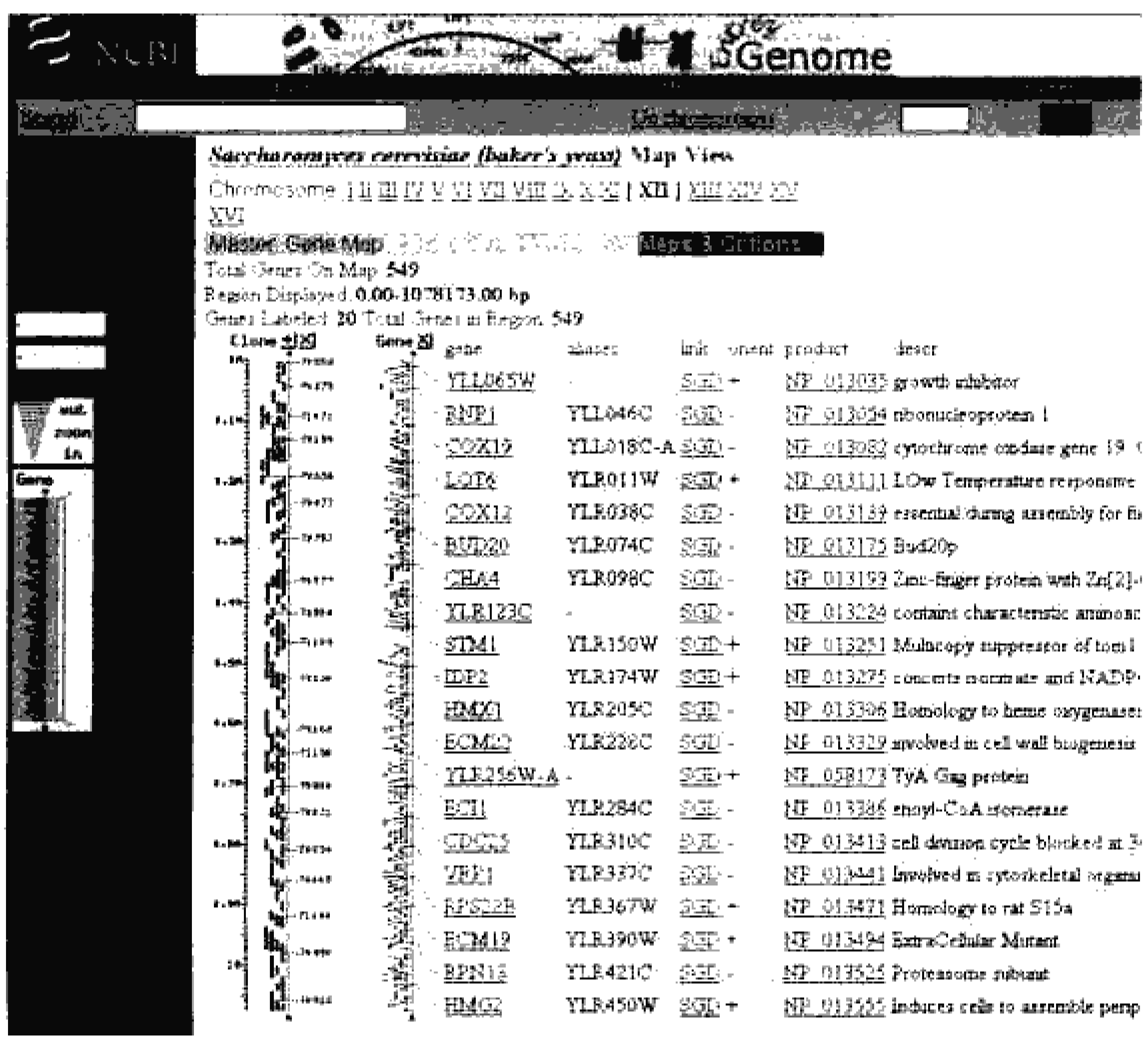


图 15-5 在边框中选择真核生物（或者“主要生物”），你能够得到关于啤酒酵母的任意一条染色体的信息。这个页面是关于第十二号染色体的，包含 COG、蛋白质结构链接；以及关于细菌和古细菌的 TaxPlot

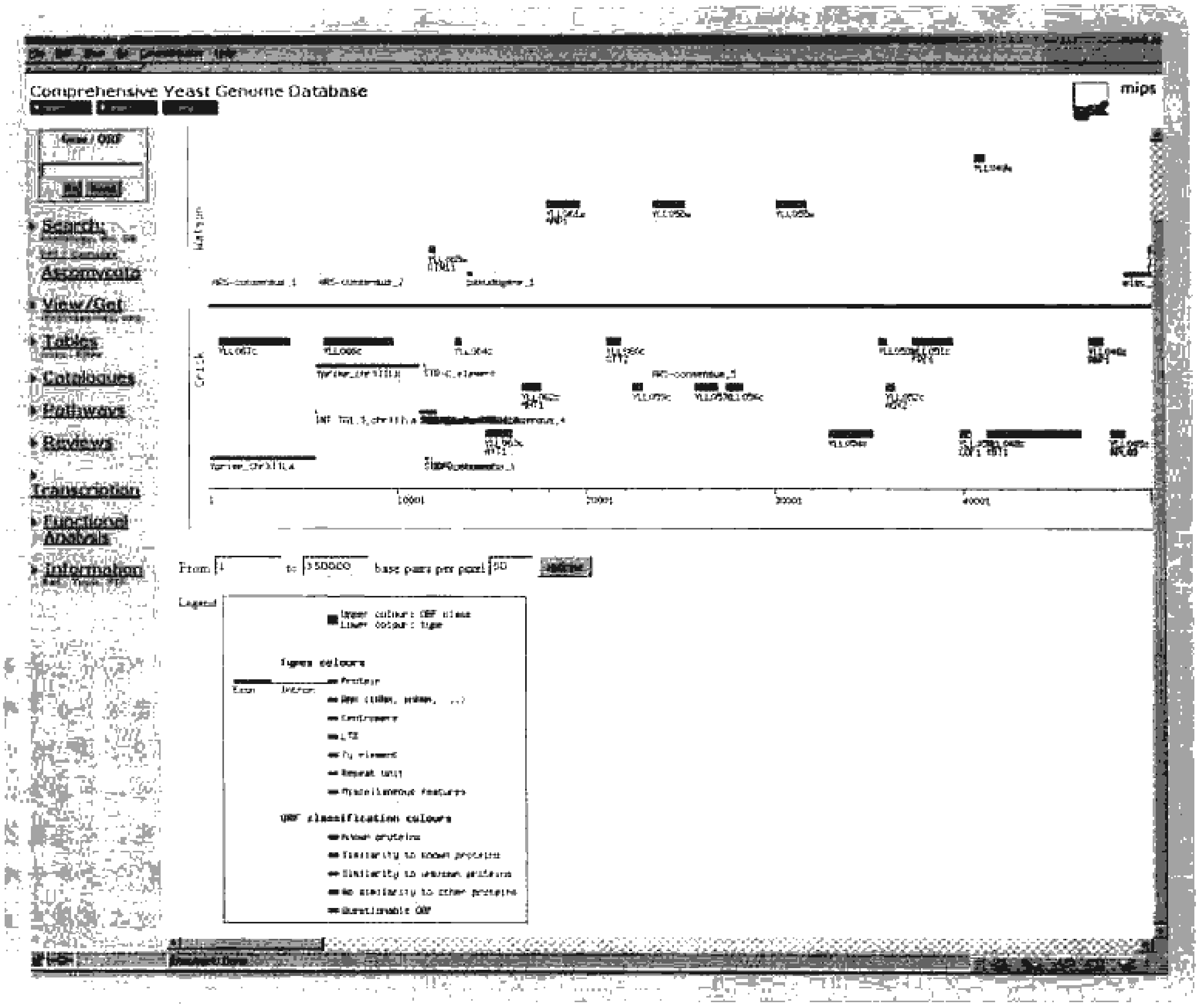


图 15-6 MIPS 提供了综合酵母基因组数据库。每个染色体能够在一张图上查看。对每个元件特征都可以点击查看详细信息。基因注释如表 15-4 所示

着丝粒是减数分裂或者有丝分裂纺锤体和染色体粘着的部分。在酵母中，着丝粒将染色体分为左臂和右臂；在人类中，它将每个染色体分为短(p)臂和长(q)臂。

端粒是每个染色体臂的末端部分。这些臂对于维持染色体结构非常重要，在衰老和智力迟钝过程（第18章）中都起着作用。

(图 15-7)。

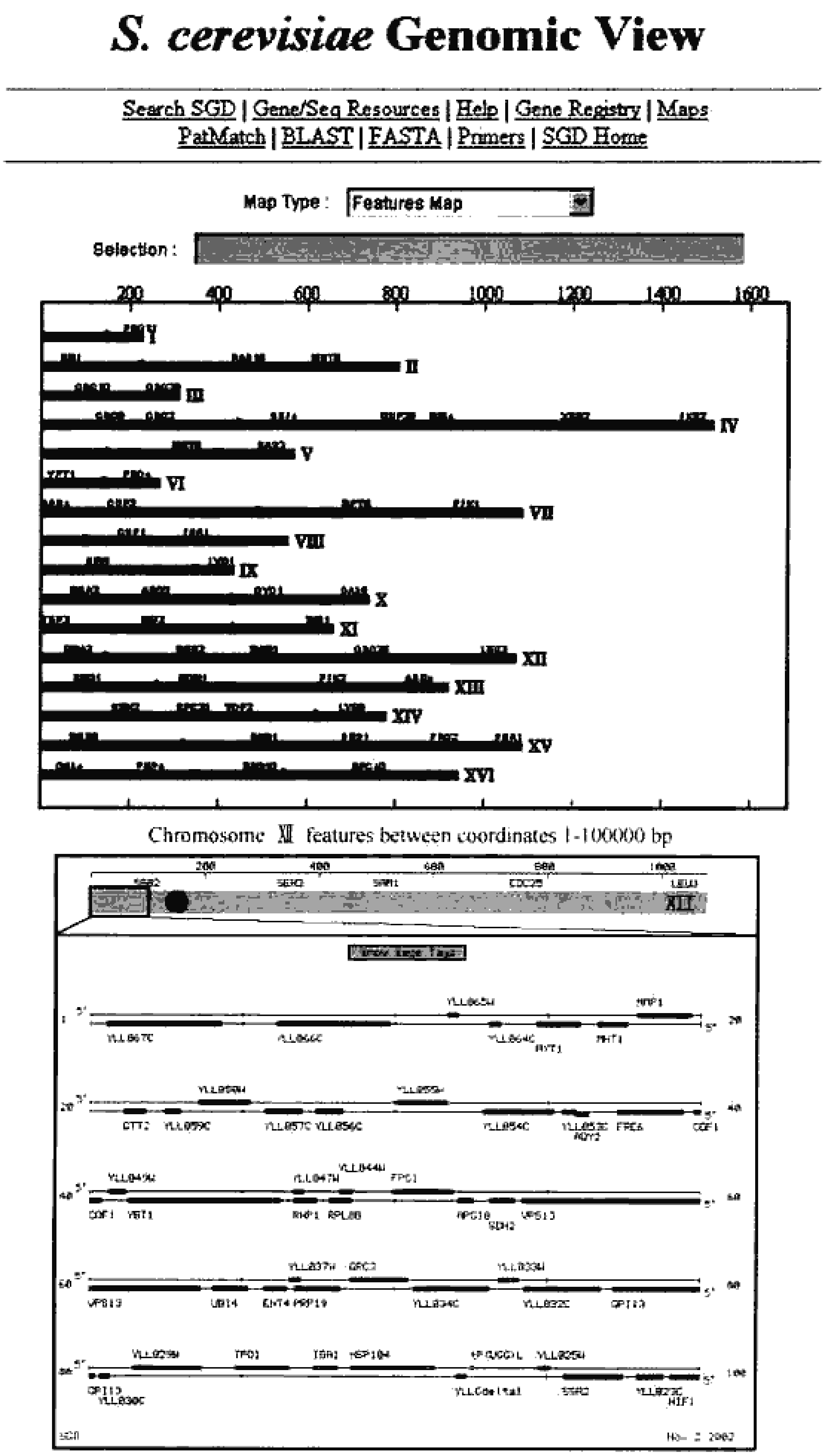
① 十二号染色体总的GC含量为38%。这个GC含量随着编码蛋白质基因的密度的提高在局部区域达到最高。有三个地方的GC含量相当低（低于37%），其中之一就是着丝粒。这是所有真核生物着丝粒都有的特征。

② 总的来说，整个啤酒酵母基因组中很少有重复DNA。在十二号染色体上有一些rDNA片段（编码rRNAs），这段区域的GC含量也是十二号染色体上最高的（大概42%）。另外，啤酒酵母染色体有端粒DNA和亚端粒DNA片段。这也是所有真核生物染色体的重要特征。

③ 剪切内含子很少（总共大概235个）。这可能是因为剪切mRNAs逆转录后产生的cDNA同源重组导致的。在十二号染色体上，17个开放阅读框（占总数的3.2%）含有内含子，其中的一般基因编码核糖体蛋白质。

④ 在十二号染色体上有6个可转座的元件（*Ty*元件），另外还有几百个可转座元件的片段。

(a)



(b)

图 15-7 (a) SGD 能够按照染色体浏览。(b) 第十二号染色体的特征为包含蛋白质编码基因、RNAs(如 rRNA、tRNA、snoRNA)转座子和 SAGE 标签

⑤ 开放阅读框的密度相当高。72%的十二号染色体中含有编码基因，这个特征也是其他酵母染色体所具有的。十二号染色体上有 534 个超过 100 个密码子的开放阅读框，这些开放阅读框的平均长度为 485 个密码子。

十二号染色体中含有啤酒酵母基因组中最大的基因——*YLR106c*。这个基因编码了一个含 4910 个氨基酸的蛋白质（MDN1p；Q12019）（Garbarino 和 Gibbons，2002）。Midasin，这个人类同源蛋白有 5596 个氨基酸（超过 600ku，NP _055426）。

啤酒酵母基因和蛋白质所采用的命名系统参见框15-2。

框15-2 啤酒酵母中的基因命名法

所有超过100个密码子的开放阅读框都被命名了一个惟一的名字，由以下几个部分组成：三个字母后面加上用来表明其在基因组上位置的数字和下标。例如，基因*YKL159c*表示在酵母（Y）六号（K）染色体的左臂（L）上的第159号开放阅读框（相对于着丝粒而言）。命名中的c或者w（"Crick"或者"Watson"）反映了这个基因在基因组中的方向。一旦一个基因被发现并具有某种功能，研究者们就会给这个基因起一个新名字来反映它的功能，如RCN1就表示"regulator of calcineurin"。显性等位基因（一般是野生性等位基因）用三个大写字母表示，隐性等位基因（一般是基因敲除突变体或者失去功能的等位基因）则用三个小写字母表示。基因的蛋白质产物不用斜体字母表示，一般是第一个字母大写并且在最后加上一个字母"p"。许多基因有多个名字（同义词），因为不同的研究者分别在独立的功能测定中发现了它们。表15-5给出了一些命名法的例子。

表15-5 酵母基因命名规则例子

野生性等位基因	蛋白质产物	突变等位基因
CNA1	Cna1p	*cna1* Δ
RCN1	Rcn1p	*rcn1*, *rcn1*::*URA3*
YKL159c	Ykl159cp	*ykl159c*

15.2.5 新基因的产生：啤酒酵母基因组的复制

随着啤酒酵母的基因组序列逐步被分析，发现有很多复制DNA序列，包括开放阅读框和大的基因组区域。很多时候，复制片段的基因顺序和方向也保留着。复制在染色体内和染色体间都有发生。

这种基因物质的改变是解释酵母物种或任何生命分支进化的基础。我们能够发现在人类基因组和一些其他真核生物基因组中，大约有25%的基因是复制产生的（第16章和第17章）。这些新复制的基因是从何处产生的呢？图15-8中列出了一些机制。

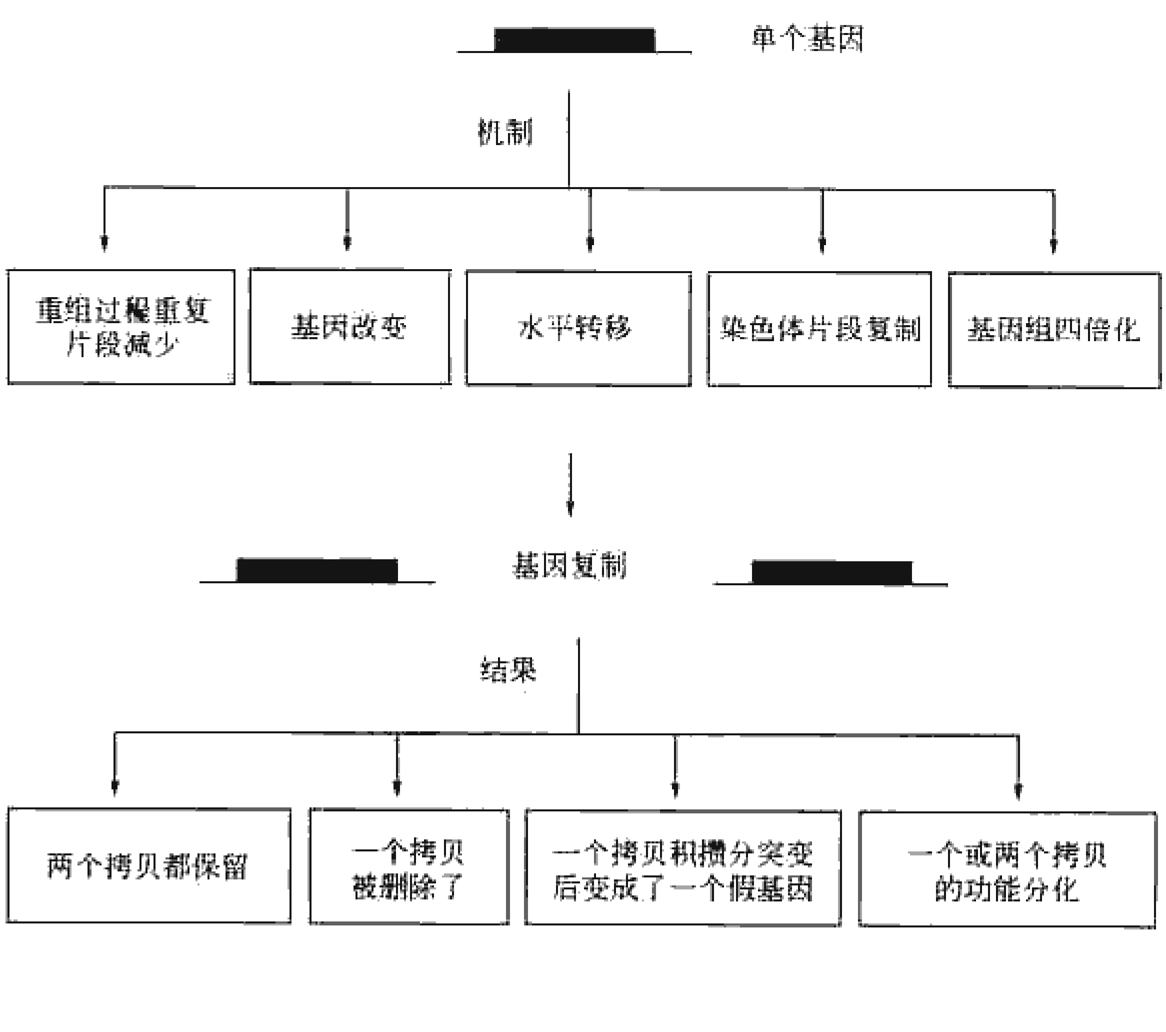


图15-8 导致基因复制的机制有多种。复制的基因可能在同一条染色体上，也可能在不同的染色体上。复制的基因有几种可能的命运：可能两个拷贝都保留，可能一个拷贝被删除，也可能一个拷贝变成无功能的（假基因），或者可能两个基因变得具有不同的功能。参见Sankoff（2001）

① 基因在复制过程中随着串联重复片段滑动而进化。

② 新基因能通过基因转移而产生。在这个过程中，基因单方向地从一个基因组区域转移到另外一个区域。这个过程主要在人类 Y 染色体的重复区域之间发生（Rozen 等，2003）。

③ 基因能够通过水平转移引入到一个基因组中。

④ 基因组的片段能够复制。

⑤ 整个基因组能够被复制，这个过程称为多倍化，在啤酒酵母中是四倍化。

> 四倍体是指在核内有四套单倍染色体

1970 年，Susumu Ohno 出版了《通过基因复制进化》。他提出脊椎动物基因组通过两轮全基因组复制进化。按照这个假说，这些复制过程发生在脊椎动物进化早期然后发展出一些不同的细胞功能。在该书的序中 Ohno 写到："要是进化过程完全依靠自然选择，那么一种细菌只能进化成为一些不同种类的细菌。也根本不可能从单细胞生物产生多细胞生物、脊椎动物以至于哺乳动物，对于进化上这么大一个跳跃，需要一些具有原来并不存在功能的基因座的产生。只有冗余的顺反子才能够逃过无情的自然选择，逃脱之后，它积累了一些原先不允许的突变成为一个新的基因座。"

啤酒酵母到底可能采用哪种基因复制机制？Wolfe 和 Shields（1997）为 Ohno 的全基因组复制假说提供了证据。他们将所有的酵母蛋白质进行系统的比对并将结果在点阵中标记出来，借此来估计酵母基因组的复制区域。复制区域在图上观察到为斜线，图 15-9 中三条斜

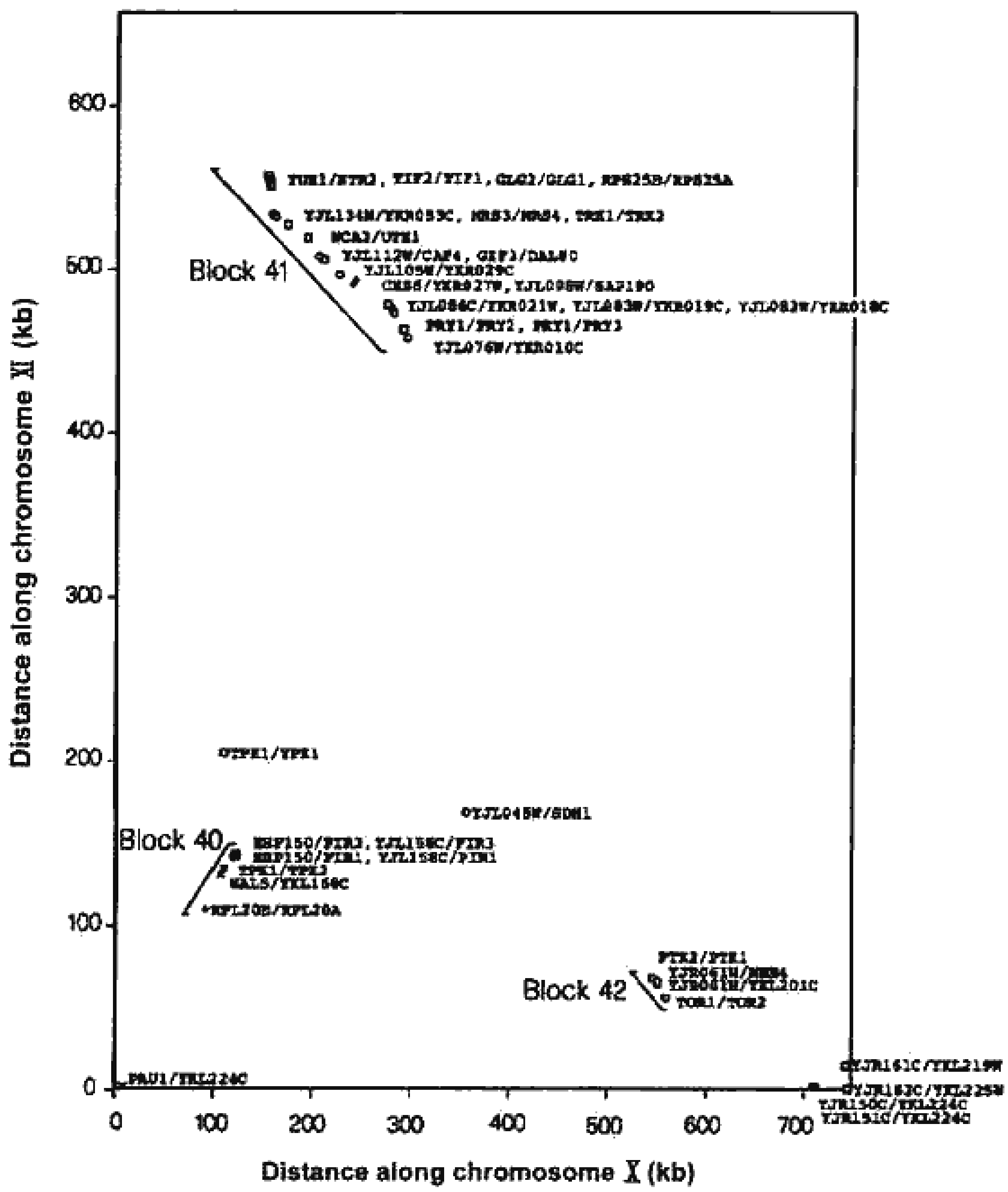


图 15-9 Wolfe 和 Shields（1997）用啤酒酵母基因组进行了 blastp 搜索，找到了 55 个复制区域。这个为整个基因组在古代曾经经历过复制提供了强有力的证据。这张图给出了第十号染色体和第十一号染色体上的基因编码的蛋白质进行 BLAST 搜索后的结果。图上显示了打分超过 200 的匹配，分成几小块排列。符号表示了基因方向：＋表示两个染色体上都是 Watson 链方向；×表示两个染色体上都是 Crick 方向；正方形和 W 表示在两个染色体上基因方向不同。授权使用

Wolfe 和 Shields (1997) 用了 Blastp 而非 Blastn 来研究染色体的复制区域。因为对于这些远源序列，蛋白质序列数据比 DNA 带有更多信息量。参见第 3 章（第 41 页到第 42 页）。

Kenneth Wolfe 和同事 (Trinity 学院，Dublin) 已经构建了用点阵对每一个啤酒酵母染色体进行比较的网站 (http://acer.gen.tcd.ie/~khwolfe/yeast/)。你在 SGD 网站上也能看到染色体之间的比较（图 15-10）

线就是酵母五号和六号染色体相比对后的复制区域。在整个基因组中，他们比对出了 55 个复制区域和 376 个同源基因。在随后的研究中，他们运用了更加敏感的 Smith-Waterman 算法又找出一些新的复制区域（Seoighe 和 Wolfe，1999）。根据这些结果，他们提出啤酒酵母基因组在古代就进行了复制（Wolfe 和 Shields，1997）。发生复制之后，许多复制的基因随之丢失了，其他一些则经单向转移后在基因组中重新排位了。

对于目前有这么多复制区域存在有两种主要的解释。一种解释为可能是转座之后的全基因组复制（四倍体）；另外一种解释为可能有一系列独立的复制。Wolfe 和 Shields（1997）更推崇那个四倍体模型，主要原因有两个。

① 这 55 个复制区域中 50 个的方向相对于着丝粒来说都是保守的。如果每个复制区域都是独立产生的，那么它们的方向应该是随机的。

② 55 个连续独立的区域复制中应该会产生大约 7 个三次复制区域，但是没有观察到（或者可能有一个）这种三次复制区域。

对 14 个半子囊菌（包括啤酒酵母和 13 个在 Genolevures 项目中部分测序的其他物种，表 15-6）的比较分析结果支持这个基因组复制后丢失并且重排的模型。

表 15-6 比较基因组学项目

生物体	染色体数目
Saccharomyces sensu stricto	
Saccharomyces bayanus var *uvarum*	16
Saccharomyces sensu lato	
Saccharomyces exiguus	14~16
Saccharomyces servazzii	7~12
Zygosaccharomyces rouxii	7
Saccharomyces kluyveri	8
Kluyveromyces	
Kluyveromyces thermotolerans	7
Kluyveromyces lactis	6
Kluyveromyces marxianus var *marxianus*	10
Distant species	
Pichia angusta	?
Debaromyces hansenii var *hansenii*	?
Pichia sorbitophila	7
Candida tropicalis	12
Yarrowia lipolytica	6

注：这个项目将 13 个近源真菌、所有的半子囊菌的基因组 DNA 进行测序。网站上（http://cbi.labri.u-bordeaux.fr/Genolevures/index.php）能输入文本进行 BLAST 搜索。

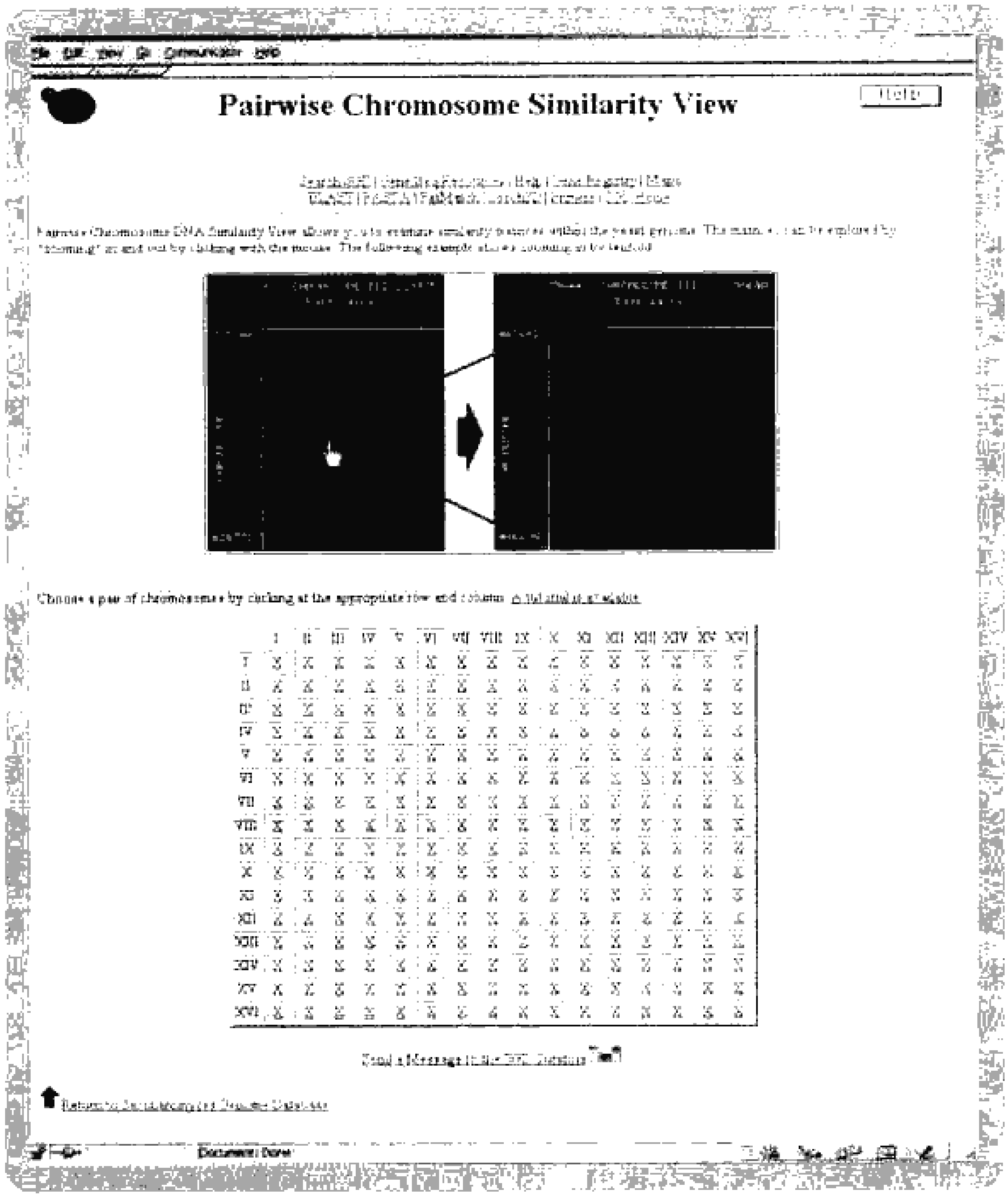


图 15-10 SGD 网站上有一个"序列分析和工具"部分，其中包含了成对染色体相似性查看器。这个工具能够让你比较两个染色体上的相关区域，显示复制区域

通过比较这 14 个生物的基因顺序，Wong 等（2002）将 70%的酵母基因组划分为只有最小交叉的姐妹区域。8 对染色体 16 个着丝粒正好与全基因组复制的模型相吻合。这些分析也表明了一些半子囊菌在全基因组复制之前就从啤酒酵母中分支出去了（如 *Kluyveromyces lactis*，*Zygosaccharomyces rouxii*，*Saccharomyces kluyveri*）。其他如 *Saccharomyces bayanus* 的半子囊菌在假设的多倍化发生后才开始分支。根据一些物种中多倍化的证据，再加上系统发生分析，Wolfe 和 Shields（1997）估计全基因组复制发生的时间大约在 1 亿年前。

Genolevures 项目在 FEBS Letters（2000 年 12 月 22 日）第 487 卷第 1 期上有一系列文章专门讨论，其网站是 http://cbi.labri.fr/Genolevures/。

啤酒酵母能够在厌氧环境下生存，但是 *K. lactis* 不能。可能是由于啤酒酵母的基因组复制产生了生理上的变化，使得它得到了这种新的生长能力（Piskur，2001）。

对于人类，如二十一号染色体多 1 个拷贝（例如，3 个二十一号染色体）就会导致 Down 综合症。十三号和十八号染色体多 1 个拷贝有时候还能够存活，但是其他常染色体不行。甚至是部分基因组的复制也会导致产生智力迟钝和其他疾病（见第 18 章有关于人类疾病部分）。这从生物个体水平突出了复制的危害性。

基因复制之后的命运如何？一般来说基因的多余拷贝对于一个生物来说是有害的。在Wolfe和他的同事的模型中，古酵母的基因组复制一遍后（从大约5000个基因的双倍体到10000个左右基因的四倍体）又失去了大部分复制基因，产生了目前的基因数目——大约6200个开放阅读框。总的来说，50%～92%的复制基因最后都丢失了（Wagner，2001）。对于真核生物，复制基因的半衰期只有几百万年（Lynch 和 Conery，2000）（参见第16章）。该过程中主要有4个可能性（图15-8）。

① 两个拷贝都留下来并且保持原有的基因功能。这种情况下，由于有了多余的基因，基因就有剂量效应。

② 其中一个拷贝丢失。这是复制基因通常的命运。对于这个可能性的合理解释为，因为最初复制的基因有相同的作用，其中之一就遭受到突变失去功能。

③ 其中一个拷贝不断积累突变进化成为一个假基因（不编码功能基因产物的基因）。这表示尽管复制的拷贝没有丢失，但是已经失去了基因功能。

④ 其中之一或者两个拷贝都改变功能。根据这种假设，基因复制（不考虑机制）能够给一个生物提供扩展功能的原材料，进而不能忍受失去具有重复功能的基因。这样功能上分开的基因就会被进化所选择。

基因复制之后，为什么新形成的基因对中之一会逐渐失活？粗略看上去，有两个拷贝是非常有益的，因为其中之一有可能在功能上分化（使得进化过程允许细胞产生新功能），或者在一个拷贝突变时另外一个还能够发挥作用。然而，基因复制一般来说是有害的，导致复制基因的丢失。这主要是因为基因中的一些突变是不允许的而非可以忍受的［这些说法是Ohno（1997）在描述基因复制时提出的］。不允许的突变严重影响基因产物的功能，例如改变一个酶的活性位点的性质（可以忍受的突变会产生一些改变，但是这些改变仍然与基因产物的功能相一致）。自然选择能够淘汰掉不允许的突变，因为这样的个体不适合繁殖下去。一个基因复制后，对于一个拷贝中的有害突变可能会忍受，因为另外一个基因能够承担它的功能。还有另外一个理由，即复制基因可能是有害的，因为它们产生后同源染色体互换时就有可能无法匹配上，从而导致不合适的互换。

我们会在图8-16中见到突触融合蛋白。

我们可以用特殊的基因例子来考虑复制基因的可能命运，这些基因编码对于囊泡输运重要的蛋白（图15-11）。在酵母和其他所有的真核生物中，球状的胞内囊泡将细胞内的各种物质转运到它们的目的地。通过将囊泡蛋白（如酵母中的Snc1p，或者哺乳动物中的VAMP/synaptobrevin）结合到靶膜蛋白（如酵母中的Sso1p和哺乳动物中的syntaxin）上，这些囊泡将物质转运到各自的靶膜上去（Protopopov等，1993；Aalto等，1993）。在啤酒酵母中，基因组复制会产生两对同源基因：*SNC1* 和 *SNC2*，还有 *SSO1* 和 *SSO2*。*SNC1* 和 *SNC2* 基因分别在一号染色体和十五号染色体的相应区域上，而 *SSO1* 和 *SSO2* 基因分别在十六号染色体和十三号染色体上。

能够在SGD上进行搜索，网址是http://genome-www.stanford.edu/Saccharomyces/。

基因组复制的后果可能是什么？这两对酵母蛋白（syntaxin-like和VAMP/synaptobrevin-like）可能会继续保持蛋白原来的功能（和基因组复制前比）。在SGD网站上搜索 *SSO1* 的结果表明该基因并不是很重要（无义突变是可存活的），但是该基因双敲除是致死的（图15-12）。这样，很可能这些同源基因为生物提供了功能上的冗余；当一个基因丢失时（如发生突变），这个生物能够依靠其他基因生存下来。同样，*SNC1* 基因无义突变后生物也是可存活的，但是 *SNC1* 和 *SNC2* 双敲除会导致分泌有缺陷。

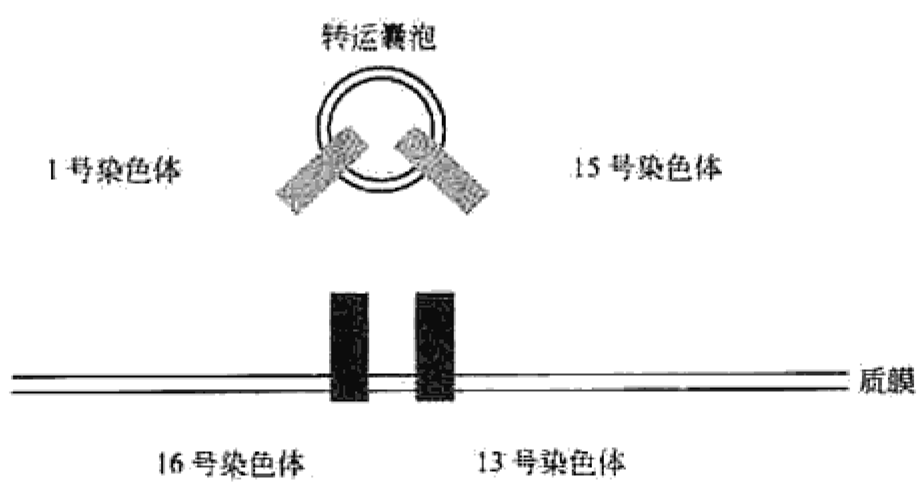


图 15-11 一对编码囊泡蛋白的啤酒酵母基因（*SNC1* 和 *SNC2*）和一对靶膜对接蛋白（SSO1 和 SSO2）。至少一个囊泡蛋白质和一个对接蛋白质结合起来作为复合物的一部分调节囊泡输运。随着整个基因组的复制（或者独立的部分基因组复制之后），同源的 *SNC1*/*SNC2* 和 *SSO1*/*SSO2* 基因含量就上升了。每一个分子中有两个拷贝就会导致功能上的冗余，所以如果其中的一个拷贝丢失了（如突变），这个生物仍然能够存活下去。或者，复制的基因能够得到不同的功能，如对接的特异性和运输的囊泡与合适的胞内靶膜的融合

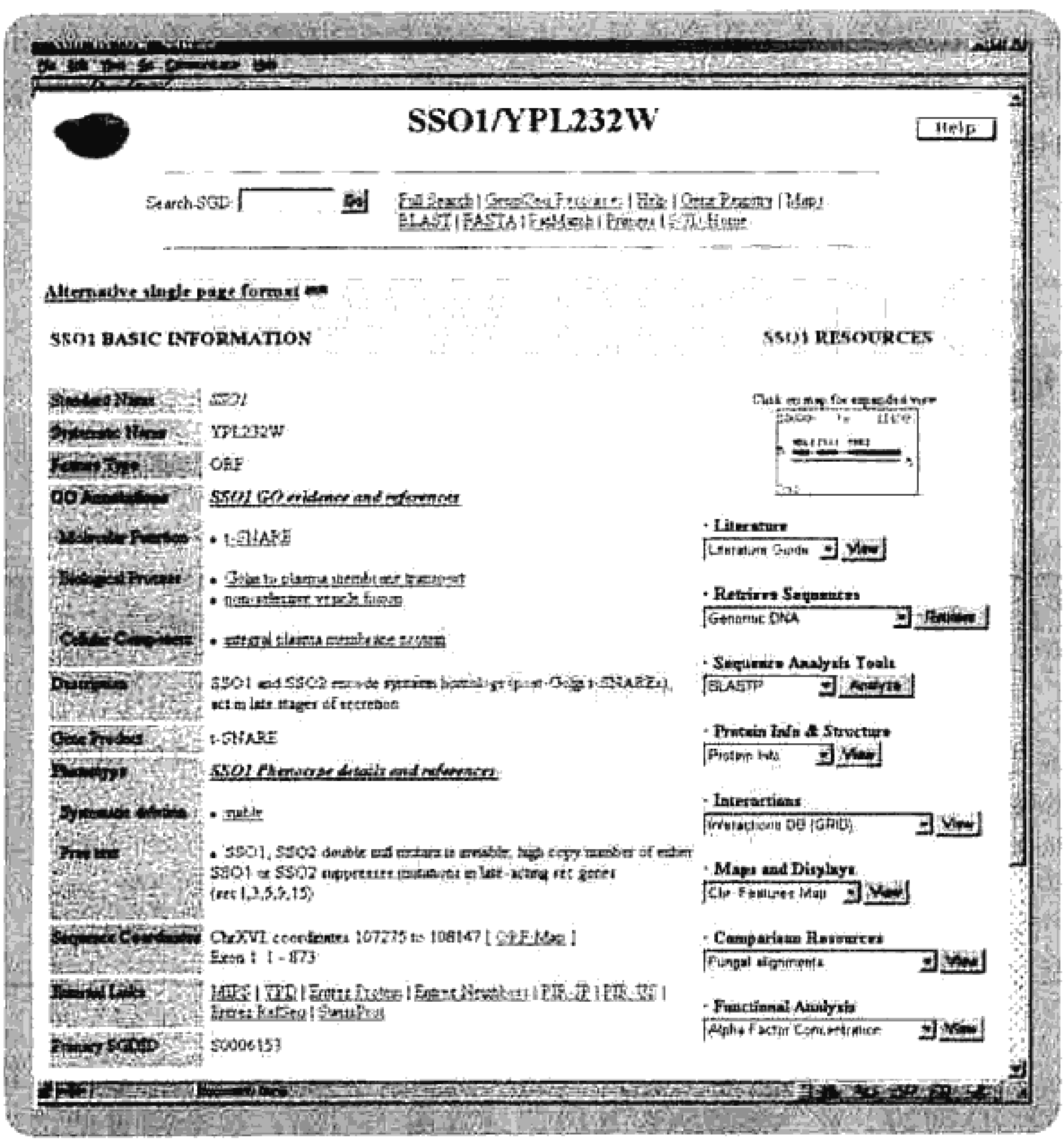


图 15-12 *SSO1* 的 SGD 记录中显示这不是一个关键基因（这个基因被删除的时候生物体是可以存活的），但是双无义突变生物体是不可存活的。SGD 的网址是 http://genome-www.stanford.edu/Saccharomyces/

还有一种对于这些基因复制的解释，即很可能全基因组复制时提供了新基因原料，而胞内分泌系统正从这些基因分化出来的。Syntaxin 和 VAMP/synaptobrevin 蛋白在不同的胞内输运步骤中起作用，并且这些基因家族在真核生物进化过程不断分化（Dack 和 Doolittle，2002）。

Andreas Wagner（2000）发表了啤酒酵母怎样通过两种机制之一来保护自己被突变的问题：①依靠有重叠功能的基因（如有相关功能的共生同源基因）或者②通过调控网络中的非同源基因的相互作用。他发现相比较失去功能后对适应性产生严重影响的基因而言，产生轻微影响的基因并没有更近的相关共生同源基因。这一点和刚才那个模型相一致，在那个模型中基因复制并不对突变具有鲁棒性。

15.3 功能基因组学计划

斯坦福生物芯片数据库是一个重要的酵母基因表达数据资源（http://www.dnachip.org/）。

功能基因组学的一种解释是指根据全基因组范围内的筛选和分析来注释基因功能。一旦一个生物的蛋白质编码基因被发现了，最大的问题就是确定它们的功能了（参见第 8 章）。传统上，结合遗传学、生物化学和细胞生物学方法，每次只能估计一个基因的功能。例如，为了确定它们可能的功能，成千上万的蛋白质被定位到细胞的亚区域，并且进行同源搜索确定可能的功能域。

在酵母基因组中，成千上万个基因的功能从基因、RNA 转录产物和蛋白质层面上（图 15-13）同时来注释。在第 8 章，我们讨论了用高通量方法来确定酵母中的蛋白质相互作用，在第 6 章和第 7 章中我们讨论了基于 SAGE 和基因芯片方法的基因表达。各种各样的网站上都有从啤酒酵母实验中得到的表达数据（表 15-7）。

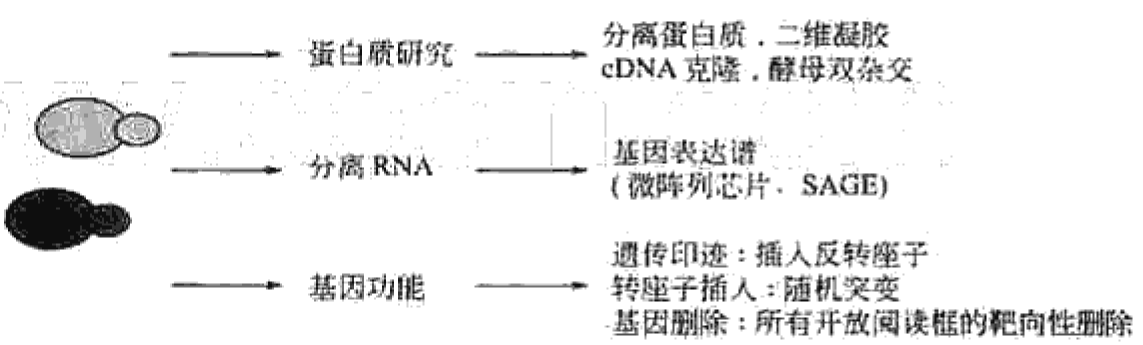


图 15-13 啤酒酵母功能基因组学方法概述。参见 Ross-Macdonald（2002）

表 15-7 万维网上的酵母表达数据

主　题	参考文献	网　址
从发酵改为呼吸的代谢改变	DeRisi 等，1997	http://cmgm.stanford.edu/pbrown/explore/index.html
孢子形成	Chu 等，1998	http://cmgm.stanford.edu/pbrown/sporulation/
细胞周期	Spellman 等，1998	http://genome-www.stanford.edu/cellcycle/
转录	Holstege 等，1998	http://web.wi.mit.edu/young/expression/
烷化剂反应	Jelinsky 和 Samson，1999	http://www.hsph.harvard.edu/geneexpression/
多药耐受	DeRisi 等，2000	http://www.biologie.ens.fr/fr/genetiqu/puces/publications/pdr1-3-3-7/
微阵列全局浏览器	Marc 等，2001	http://www.transcriptome.ens.fr/ymgv/

注：来源于 MIPS 数据库（http://mips.gsf.de/proj/yeast/CYGD/db/index.html，2002 年 11 月）。

在本章中我们将会讨论其他一些通过大规模基因筛选后高通量测定基因功能的方法。作为一个起始点，我们先在 SGD 网站上看一下 *SSO1* 基因记录（图 15-12）。这个网页的边框上有一些下拉菜单，通过这些菜单可以很方便地看到一些功能基因组的信息，包括蛋白质相

互作用数据（图 15-14 和图 15-15），在不同条件下基因表达的改变（图 15-16），还有一些其他的分析。另外还有一些结构信息，如酵母基因和其他真菌相应的基因组 DNA 之间的关系（图 15-17）。

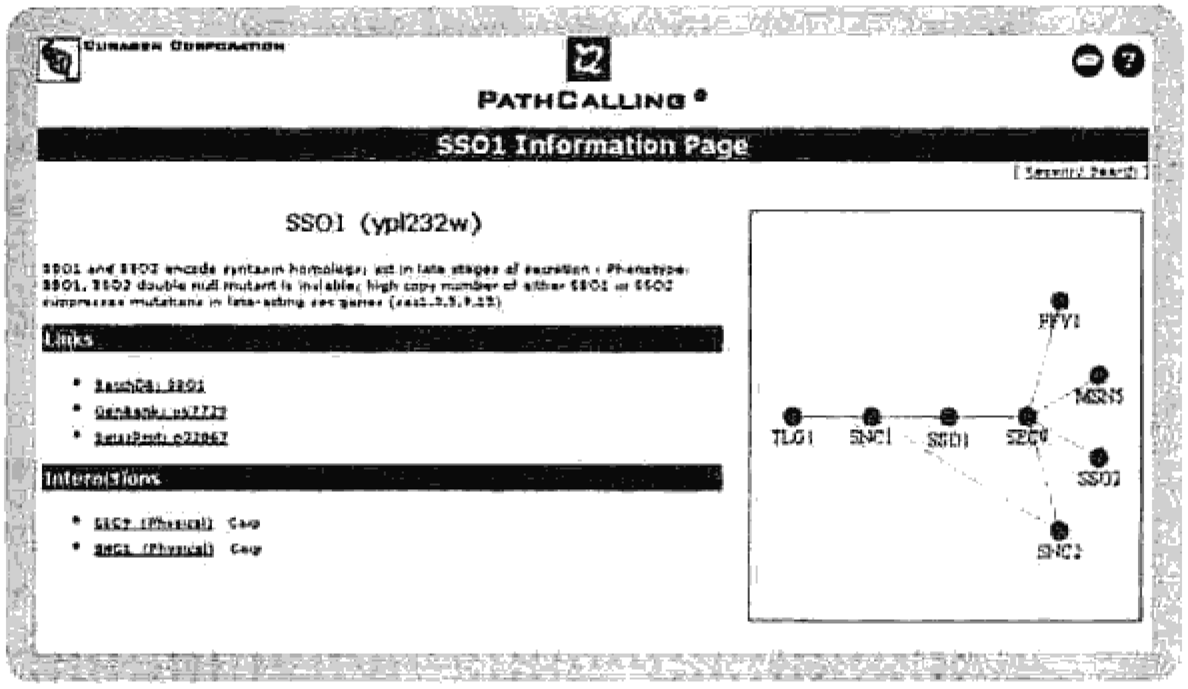


图 15-14　SGD 提供了一个链接，输入一个典型的基因（*SSO1*），就能够链接到基于酵母双杂交（第 8 章）的已知相互作用。Ssolp 在异复合体中结合到其他蛋白质上一起调节囊泡的对接和融合。所有图上出现的基因在人类中都有同源基因，在释放神经递质的突触囊泡的对接和融合过程中发挥作用。这就更加突出了啤酒酵母作为一个模式生物在真核生物生物学过程的研究中的作用

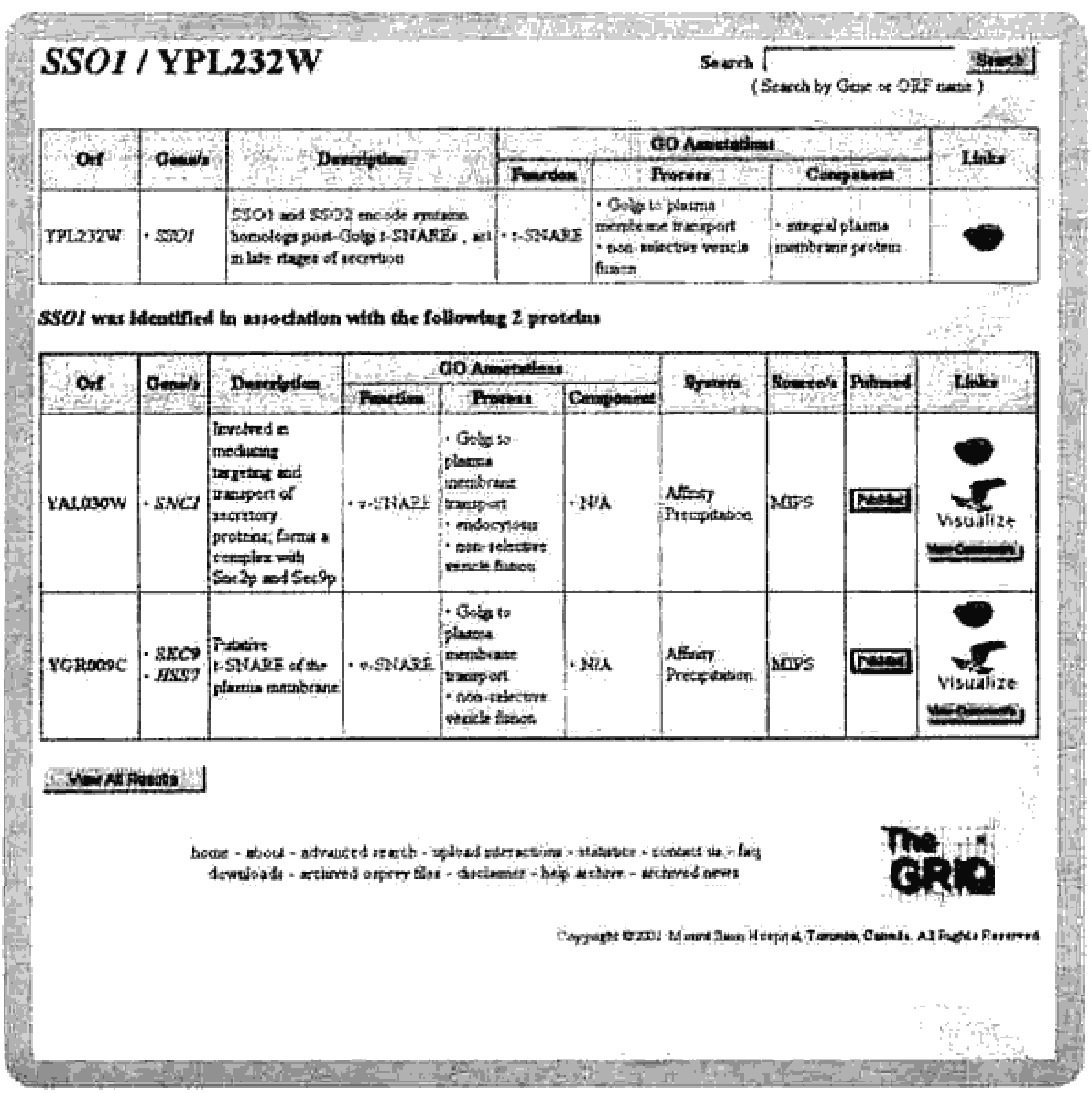


图 15-15　在 SGD 中输入 *SSO1* 后链接到了 GRID 数据库，这个数据库显示了和 Ssolp 有相互作用的蛋白质。GRID 的网址是 http://biodata.mshri.on.ca/grid/servlet/Index

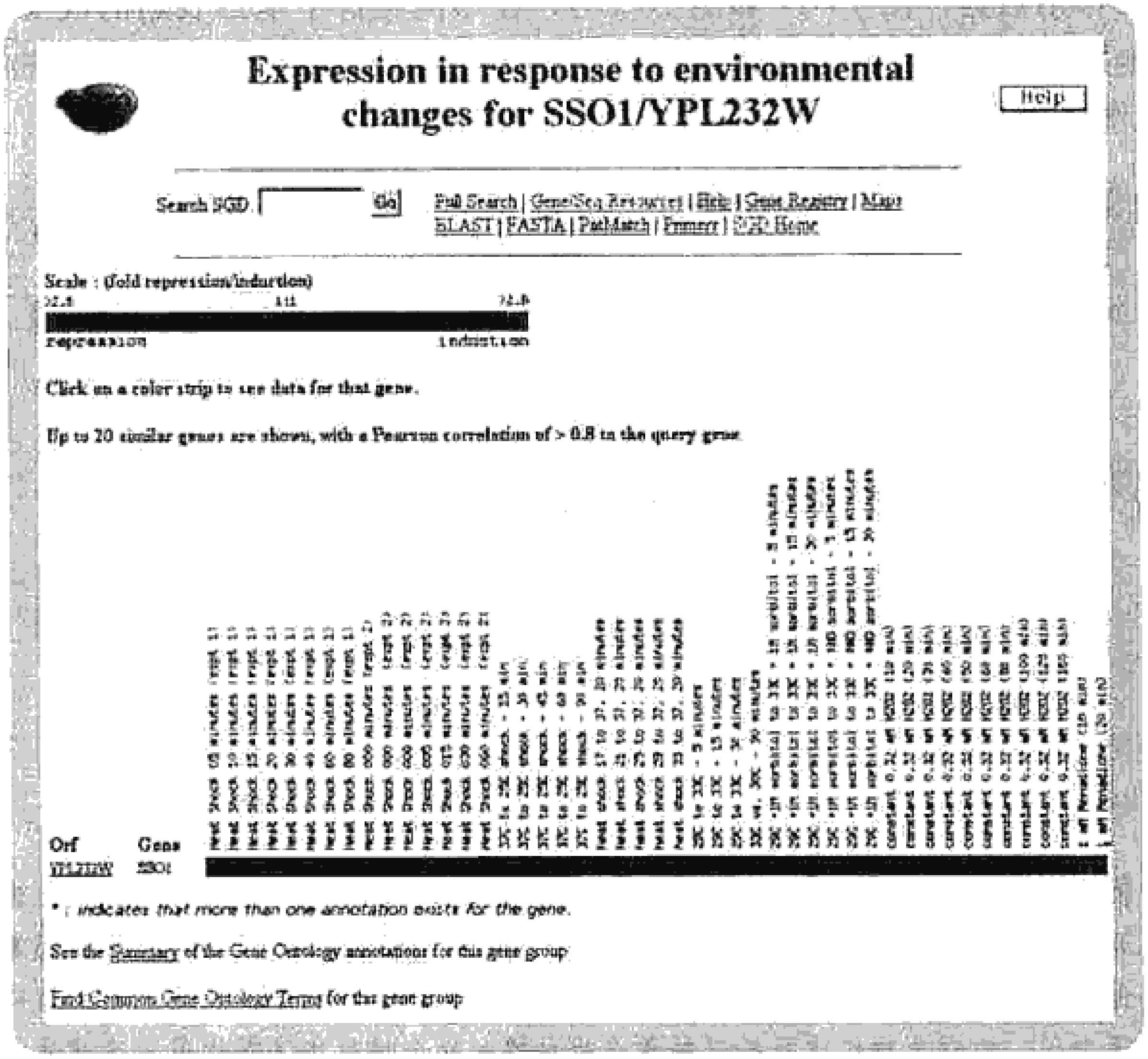


图 15-16 SGD 中关于一个典型的基因（*SSO1*）的资料表明它的表达水平随着不同的环境变化而改变

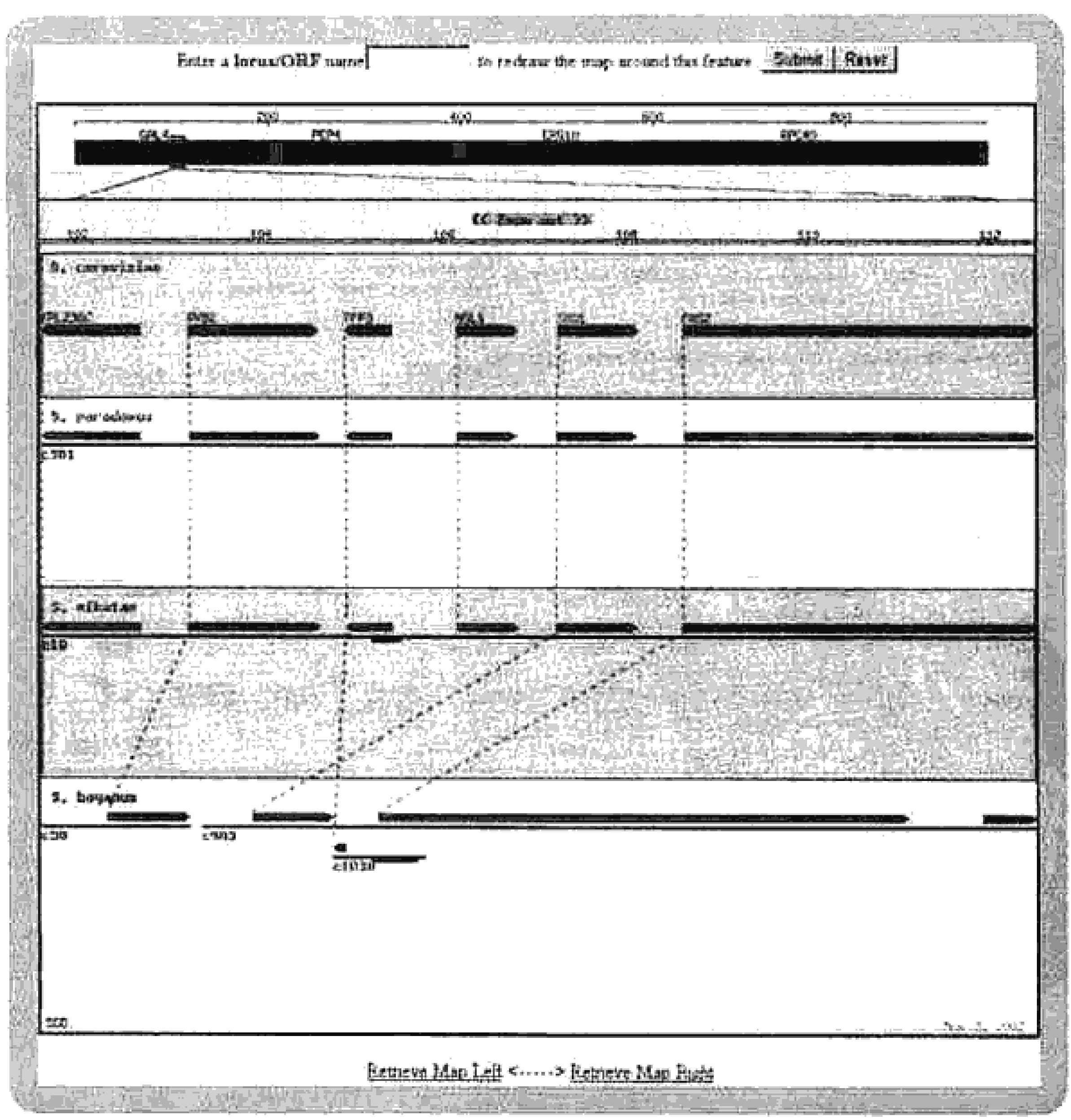


图 15-17 SGD 上的synteny 查看器显示了十六号染色体上的基因组 DNA 片段（*SSO1* 基因所在的地方）和相关的真菌基因组上相应的 DNA 区域

我们接下来还会讨论一些大规模考察基因功能的方法。如有了全基因组，就能对所有已知的酵母基因进行全面的功能分析。

15.3.1 用转座子进行基因印迹

Ty1 元件是一个酵母逆转座子，它随意地插入在基因组中。Patrick Brown、David Botstein 和他们的同事们发明了一种方法，使得酵母在几个不同条件下生长（如丰富的培养基和最小培养基）并且受到 *Ty1* 转座子调控的突变（Smith 等，1995，1996）（图 15-18）。插入之后，聚合酶链反应（PCR）采用基因和 *Ty1* 元件特定的引物。这会导致产生一系列不同大小的 DNA 产物。这个方法的前提是单个基因（如 *SSO1*）在一定条件对生长非常重要，所以 PCR 产物就会有损失（基因印迹），从而表明该基因在特定条件下的重要性。

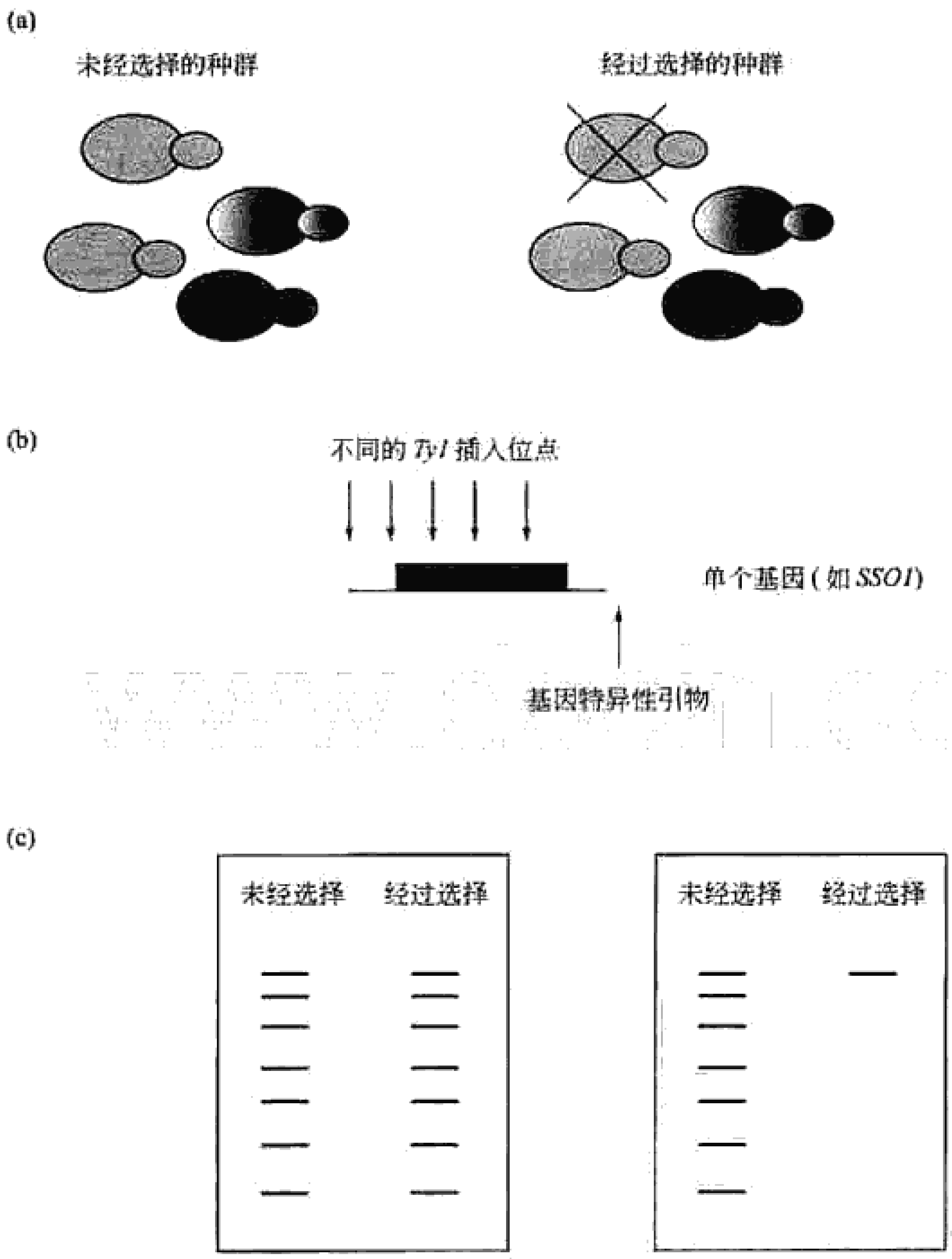


图 15-18 基因印迹：(a) 选择一个酵母种群，例如改变培养基或者加入一些药物。有些基因不会被选择过程所影响；(b) 随意插入一个转座子，进行基因特异性 PCR；(c) 接下来观察凝胶电泳后的 DNA 产物。有些基因不会被选择过程所影响（左）。另外一些由转座子标记的基因将会由于体积减小而减少，从而 PCR 产物就会减少，这样就能够证明这个基因是酵母在这个选择条件下生存所必需的

这个方法有以下优点：

① 任何感兴趣的基因都能拿来分析或者说能够随意选择基因。

② 任何给定的基因都能进行多位点突变分析。

③ 可以在一个群体中同时进行表现型分析。

④ 能够选择许多不同的表现型进行分析。

⑤ 对于重叠基因这个方法也适用。

但是也有一些缺点：

① 突变株无法复原。

② 能进行多位点突变（等位基因），所产生的都是插入突变（不是基因敲除或者其他突变）。

③ 这个方法劳动强度很大，需要一个基因一个基因地进行分析。

④ 具有重叠功能的复制基因的作用可能会被忽视掉。

15.3.2 开发外源转座子

另外一种基因突变方法是用细菌活酵母转座子作为标记物随意插入基因中去（Ross Macdonald 等，1999）（图 15-19）。有一个迷你转座子来自一个细菌转座子 *Tn3*，其中有一个 *lacZ* 报告基因，没有启动子甲硫氨酸或者上游启动子序列。当它随意地插入到一个蛋白质编码基因中后，有 1/6 的可能性会按照开放读码顺序进行翻译。一旦发生这种情况，酵母就会产生 β-半乳糖苷酶来发现这个插入转座子。这个模型中包括 *loxP* 位点，使得产生基因重组。在重组过程中，这个基因被标记上一小段 DNA 片段，这个片段编码了 3 个红细胞凝集素抗原决定基标签拷贝。

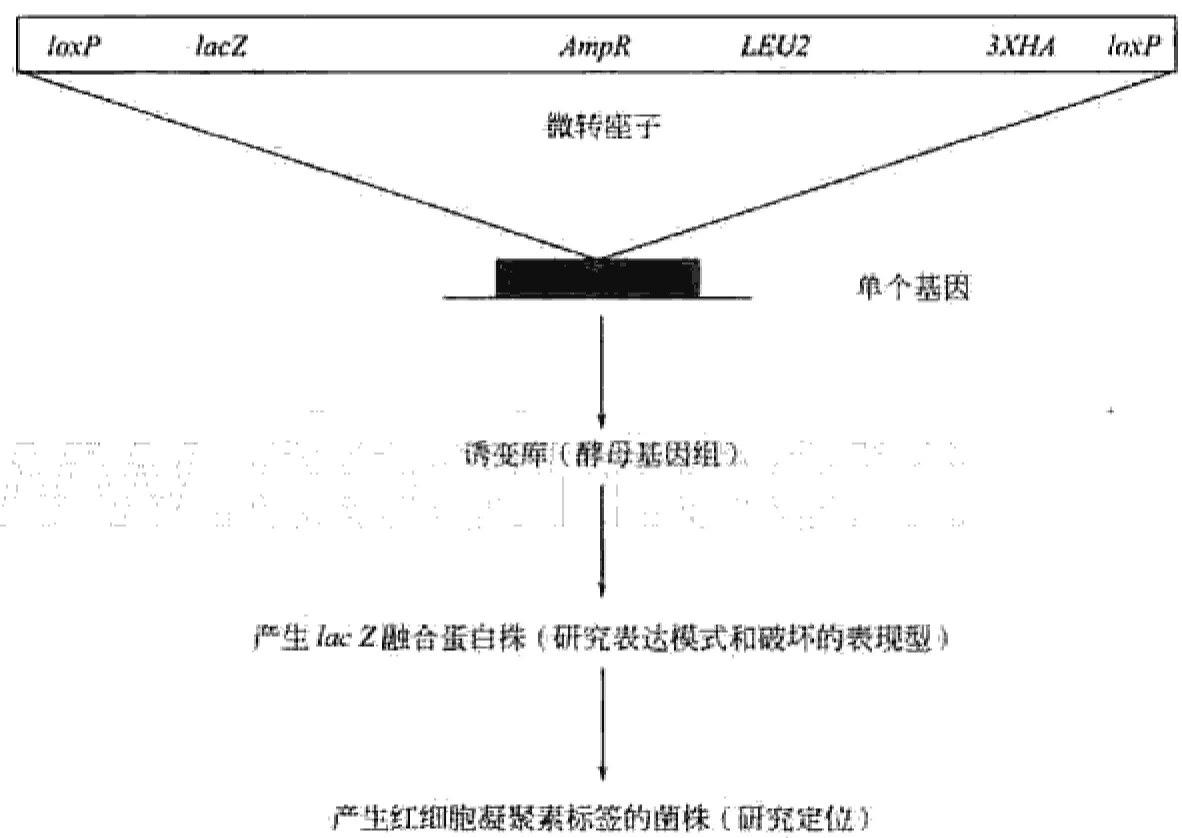


图 15-19 通过转座子标签和基因破坏来研究酵母的基因功能（改编自 Ross-Macdonald 等，1999）

> 我们在第 8 章讨论了 HA 标签。一个带上 HA 标签的蛋白质能够用 HA 特异性的抗体在细胞内进行定位。
> TRIPLES 数据库的网址是 http://ygac.med.yale.edu/triples/triples.htm。TRIPLES 是酵母中转座子的插入、表现型和表达的首字母缩写。在 1999 年，这个数据库中有超过 92000 个质粒片段、11000 个蓝色酵母菌株（对 β-半乳糖苷阳性）、6300 个已经测序的酵母菌株、1900 个已知开放阅读框和 1346 个和已知开放阅读框融合的部分。

这种迷你转座子能够进行全基因组范围的干扰表现型、基因表达和蛋白质位点分析。Ross Macdonald 等（1999）用 11000 个酵母株来进行 20 个不同条件下的干扰表现型分析。他们的研究发现了 300 个以前没有进行注释的开放阅读框。这个实验中的数据被存放在 TRIPLES 数据库中（Kumar 等，2002b）。图 15-20 是在这个数据库中查询 *SSO2* 的结果。

(a)

Query Results   Help

The search conditions are: Gene Name: sso2
The output results are sorted by: Clone ID

Clone ID	Gene or ORF (plus synonyms)	mTn insertion	Insertion point	Total protein length
TN7-113D8	SSO2/YM8010.13/YMR183C	in-frame	94	295
TN7-95G9	SSO2/YM8010.13/YMR183C	out-of-frame	157	295
V195A4	SSO2/YM8010.13/YMR183C	in-frame	272	295
V10G6	SSO2/YM8010.13/YMR183C	out-of-frame	-116	0

- **Please click on the highlighted clone ID to obtain a composite report of phenotypic, expression, and protein localization data for a given clone.**
- **Please access our on-line help for a more complete description of the data presented above.**

(b)

V195A4   Request reagents

Genbank accession	SGD Link	YPD Link	Genecensus Link (Gerstein's Lab)
none	YMR183C	YMR183C	YMR183C

Potential ORFs Disrupted by mTn Insertion   Help

Gene or ORF (plus synonyms)	mTn insertion/ORF features	Chr.	Chr. Insertion Coord.	Insertion point (codon #)	Total ORF length (amino acids)	Repeated Gene	LacZ Orientation
SSO2/YM8010.13/YMR183C	In frame	XIII	626998	272	295	may be	sense

Gene Expression Data   Help

Growth condition	LacZ expression level	Background strain
vegetative	intense blue	Y800
sporulation	intense blue	Y800

Subcellular Localization   Help

Localization	Cell stage	% of cells staining	Stain intensity	Trials	Comments
cyto. (patches) / Endo. Reticulum	all	35	2	2	patchy staining of the cytoplasm and ER

图 15-20 (a) 一个在 TRIPLES 数据库中搜索 *SSO2* 的结果的例子；(b) 转座子插入后的突变体及其表现型

这种大规模方法能够得到许多信息：

① 令人惊奇的是，480 个表达的插入序列在同一个开放阅读框中，但是顺序不正确。这就说明阅读框平移可能是一个相当常见的基因表达机制。

② 有 328 个转座子按照读码顺序插入到了没有注释的开放阅读框中（这些阅读框长度从 50 个密码子到 247 个不等）。这样，这个方法就能够用来找到一些新的蛋白质编码基因。

③ 在发现的这些以前没有注释的开放阅读框中，52%是已知开放阅读框的反义框，15%和另外一个框中的已知开放阅读框重复，剩下的 37%则是独立基因。这些发现正好与还有很多小基因没有被发现的假设所吻合。

这个方法也有许多有用的特点：

① 它包含了基因表达水平和蛋白质位点的数据。

② 就像前面所叙述的，能够发现新基因。

③ 对于重叠基因这种分析也有效。

④ 突变株能够恢复和存放。它们能够通过 TRIPLES 数据库被不同的科研单位使用。

这个方法也存在一些问题，即和前面介绍的基因印迹法差不多，需要转座子位点特异性，失去了一些具有复制功能的基因信息，并且也是一项高劳动强度的工作。

15.3.3 根据分子条形码进行全基因组范围的删除

有一个研究组织完成了一个非常的实验，就是能够培养出删除任意一个已知基因的酵母突变株（Winzeler 等，1999；Giaever 等，2002）。这个实验项目的目标如下：

出芽酵母有两种交配形式：a 和 α。单倍体 a 和 α 细胞能够相互交配形成双倍体 a/α 细胞。在细胞周期过程中单倍体和双倍体都靠有丝分裂来生长。

① 培养出一个酵母基因敲除库，所有的啤酒酵母基因组中大约 6000 个开放阅读框都被敲除过。

② 培养所有非必需基因（大约占总基因的 85%）的几种有用形态：基因敲除的杂合二倍体，基因敲除的纯合二倍体，接合型和 α-接合型酵母株。

③ 培养所有的必需基因（大约占总基因的 15%）敲除的杂合二倍体。

这个项目采用通过 PCR 进行基因替代的方法，主要依赖于酵母中高频率的同源重组（图 15-21）。根据每个基因的上游或者下游区域，一小段 DNA 片段（大概 50 个碱基对）被

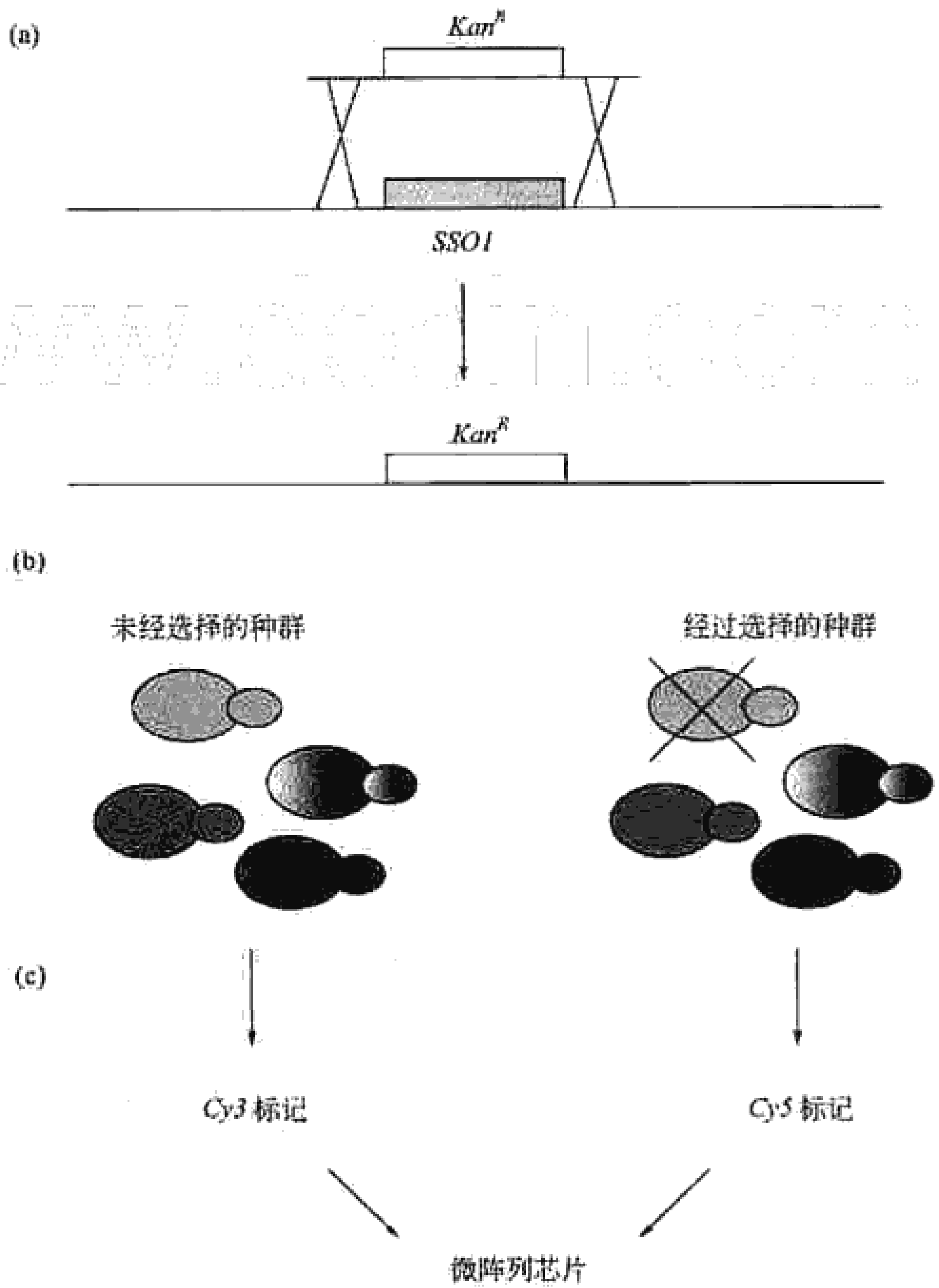


图 15-21 所有啤酒酵母基因的靶向性删除。(a) 这个方法是通过同源重组后进行基因替代。每个基因删除后用 Kan^R 基因替代，上下游都有特异的标签引物序列；(b) 使用不同的选择条件；(c) 在每种条件下分离出基因组 DNA，标记上 *Cy3* 和 *Cy5*，然后和基因芯片进行杂交，用这种方法，在每种生长环境发挥功能的基因就能够鉴别出来了

插入到一个挑选出来的标记基因后面。另外，两个“分子条形码”——一个上游标签和一个下游标签是惟一的二十个碱基对的寡核苷酸序列，它们被插入到每一个删除株中去。这个特点使得成千上万个删除株可以同时在不同的生长条件下被检测。Giaever 等（2002）记录了在 6 种条件下最优生长所必需的基因，这 6 种条件分别为高盐、山梨糖醇、半乳糖、pH8、小量培养基并且加入了抗真菌药制霉菌素。在这些发现中：

① 大约 19%的酵母基因（1105 个）在大量葡萄糖培养基中生长是必需的。这些基因中只有大约一半是早已知道是必需的。

② 非必需开放阅读框更可能编码新蛋白质。

③ 必需基因更可能在其他生物中有同源基因。

在酵母基因组中很少有必需基因复制（8.5%的非必需基因有共生同源基因，而只有 1%的必需基因有共生同源基因）。这一点就支持了复制基因有重要的冗余功能这个假说（参见第 517～第 518 页）。

这个系统基因删除方法有许多重要的优点：

① 所有在啤酒酵母中已知的基因都可以用来分析。

② 每一个突变都被定义了同一个结构。

③ 能够保证突变是无效的。

④ 基因敲除突变株都能够复原、储存并能够供其他科研团体使用。

⑤ 很容易进行多基因家族的研究。

⑥ 使得并行表现性分析成为可能，能研究许多不同的表现型。

⑦ 一旦培养出一个突变株，在研究其表现型时只需要很少的劳力投入。

这个方法也有一些缺陷：

① 培养这些基因敲除突变株需要很大的劳力投入。

② 对于每个基因，只能培养出无效等位基因进行研究（另外的等位基因可能从其他研究中得到）。

③ 相比较转座子随意插入方法，这个方法没有发现新基因。

④ 所有没有注释过的开放阅读框都被忽略了，尤其是那些短开放阅读框很可能就不能被注释了（参见上文）。

⑤ 很难解释重叠基因的删除。

15.4 真菌基因组分析

除了啤酒酵母，现在很多其他真菌基因组也都完成了测序（表 15-8）。我们将会按照字母表顺序来进行讨论。裂殖酵母（*S. pombe*）基因组是第二个即将被完全测序的真菌基因组，再接下来就是 *Neurospora crassa* 基因组。

表 15-8 真菌基因组测序项目

生物体	大小/Mb	染色体数目	网址
Aspergillus fumigatus	30	8	http://www.sanger.ac.uk/Projects/A_fumigatus/ http://bioinf.man.ac.uk/ http://www.pasteur.fr/externe http://www.tigr.org/

续表

生物体	大小/Mb	染色体数目	网址
Aspergillus nidulans	28.5	8	http://www.cereon.com/
Aspergillus parasiticus	NA	NA	http://www.genome.ou.edu/fungal.html
Candida albicans	16	8	http://sequence-www.stanford.edu/group/candida/
Cryptococcus neoformans	21	NA	http://reweb.bcgsc.bc.ca/cgi-bin/cryptococcus/cn.pl http://www.nagasaki-u.ac.jp/english/ http://www-sequence.stanford.edu/group/C.neoformans/ http://www.genome.ou.edu/cneo.html
Fusarium sporotrichioides	NA	NA	http://www.genome.ou.edu/fsporo.html
Magnaporthe grisea	40	7	http://www.wi.mit.edu/ http://www.fungalgenomics.ncsu.edu/index.htm
Neurospora crassa	43	7	http://www-genome.wi.mit.edu/ http://www.unm.edu/~ngp/ http://www.genome.ou.edu/fungal.html
Phanerochaete chrysosporium	30	10	http://www.jgi.doe.gov/programs/whiterot.htm
Saccharomyces cerevisiae	12	16	http://genome-www.stanford.edu/Saccharomyces/
Schizosaccharomyces pombe	13.8	3	*Nature*, vol. 415, pp. 871-880, 2002
Ustilago maydis	20	NA	LION Bioscience and Bayer

注：1. 来源于 NCBI。
2. NA 表示未获得。

15.4.1 烟曲霉

TIGR 网站上提供了关于 *A. fumigatus* 的概述和目前其基因组测序的进度（http://www.tigr.Org/tdb/e2k1/afu1/）。*Aspergillus* 的网站（http://www.aspergillus.man.ac.uk/）上有详细信息。

烟曲霉（*Aspergillus fumigatus*）是世界上最为常见的导致感染的霉，它被划分为丝状菌。其基因组测序是由下面几个单位完成的：基因组研究所、桑格研究中心和巴斯德研究所。

15.4.2 白念珠菌

C. albicans 的基因组是由斯坦福大学进行测序的（http://sequence-www.stanford.edu/group/candida/index.html）。关于这个真菌的更多信息可以在 http://alces.med.umn.edu/Candida.html 上得到。

白念珠菌（*Candida albicans*）是一种双倍体有性真菌，它能够导致人类偶然性的感染，其基因组大小大概是 16Mb。它的染色体和其他染色体不一样：基因组有 8 对染色体，其中 7 对是稳定的，剩下来的 1 对是可变的（在 3～4Mb 变化）。

15.4.3 典型的真菌：小孢子寄生虫

小孢子虫（*Encephalitozoon cuniculi*）是没有线粒体和过氧物酶体的单细胞真核生物。它们作为必需的胞内寄生虫感染动物（包括人类）。小孢子虫 *E. cuniculi* 的全基因组由法国的几个研究小组测定（Katinka 等，2001）。它的基因组相当紧凑，2.9Mb 中有 2000 个蛋白质编码基因。与寄生细菌（第 14 章）相似，这些病原体的基因组大大减小。对一些 *E. cuniculi* 蛋白进行系统发生分析表明，这些寄生虫不是典型的真菌，它们曾经有线粒体，但是随后就消失了（图 15-22）（Katinka 等，2001）。

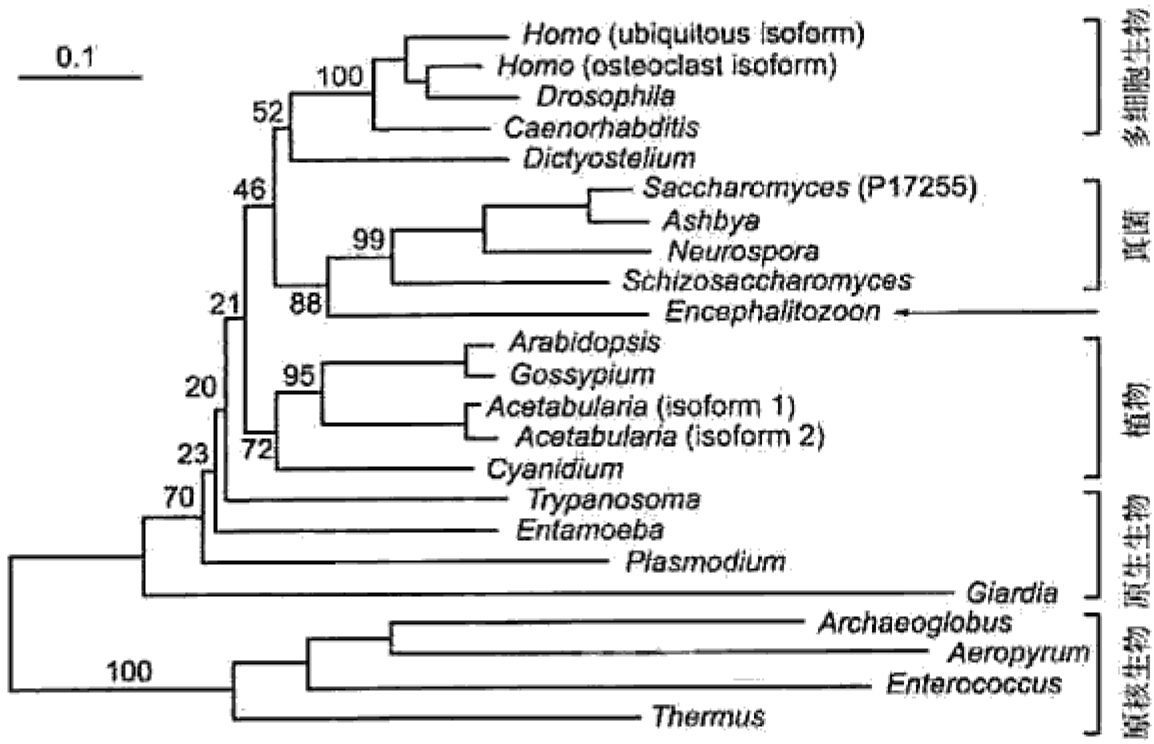


图 15-22 对动物、植物、真菌、原生生物和原核生物的液泡 ATP 酶亚基 A 的系统发生分析结果支持小孢子寄生虫 *Encephalitozoon cuniculi*（箭头处）是从真菌起源的。这个发生树是 Katinka 等（2001）用邻接法得到的。图中值是引导百分数（参见第 11 章）。授权使用

15.4.4 脉孢菌

自从 Beadle 和 Tatum 在 20 世纪 40 年代利用橘子面包霉建立了一个基因-一个酶模型后，橘子面包霉脉孢菌（*Neurospora crassa*）就一直作为一种既漂亮又简单的模式生物用于基因和生化研究。脉孢菌是丝状菌中研究得最为透彻的，丝状菌对于农业、医药和环境都相当重要（Perkins 和 Davis，2000）。脉孢菌进化上的复杂性和其他单细胞酵母差不多（Casselton 和 Zolan，2002）。就像果蝇一样，脉孢菌在自然界中广泛存在，所以它很适合于大规模研究。

> 脉孢菌属基因组数据库能在下面这些网站上得到：新墨西哥大学（http://www.unm.edu/~ngp/），Whitehead 研究所（http://www-genome.wi.mit.edu/annotation/fungi/neurospora/）和 MIPS（http://www.mips.biochem.mpg.de/proj/neurospora/）。
>
> George Beadle 和 Edward Tatum 分享了 1958 年的诺贝尔奖（和 Joshua Lederberg），因为他们发现了基因调控一定的化学事件（http://www.nobel.se/medicine/laureates/1958/）。他们用 X 光对 N. crassa 进行照射来研究基因功能。

和啤酒酵母一样，脉孢菌是子囊菌，因此就具有这种生物的优点，可以在实验室产生完整的四分体用于基因研究。然而，在很多重要方面比起酵母来它都更像动物。例如，它在呼吸链中有复合体Ⅰ，有明显可区分的昼夜节律，并且它用 DNA 甲基化来控制基因表达。60 多年对脉孢菌遗传学、生物化学和细胞生物学的广泛研究证明了这种生物是研究和发现生物学知识的金矿。

Galagan 等（2003）报道了脉孢菌的全基因组序列。他们对 7 个染色体上的大约 39Mb DNA 进行了测序，并且鉴定出了 10082 个蛋白质编码基因（9200 个长于 100 个氨基酸）。这些蛋白质中，41％没有相似的已知序列，57％在啤酒酵母或裂殖酵母中没有可识别的同源蛋白。

脉孢菌基因组只有 10％的重复 DNA，包括大概 185 个 rDNA 基因拷贝（Krumlauf 和 Marzluf，1980）。其他重复 DNA 都是分散的并且都比较短和/或者分化了。这很可能是因为“RIP”现象（重复诱导的点突变，repeat-induced point mutation）。RIP 就是在特殊的减数分裂前期细胞的单倍体细胞核中查看基因组复制（重复）序列的机制。RIP 机

制能够很有效地找到重复序列并使得它们发生一些GC向AT的突变（Selker，1990）。很明显RIP是作为脉孢菌的一种基因组防御系统，使转座子失活并防止基因组扩增（Kinsey等，1994）。Galagan等（2003）发现仅有很少脉孢菌基因属于多基因家族，还发现了仅有8对复制后的基因编码超过100个氨基酸的蛋白质。另外，81%的重复DNA序列都通过RIP突变了。这样，RIP就通过在脉孢菌中复制抑制了新基因的产生（Galagan等，2003；Perkins等，2001）。

15.4.5 第一个担子菌：白腐真菌

白腐真菌（*Phanerochaete chrysosporium*）是第一个全基因组测序的担子菌门的真菌。它是一种分解其他物质（包括污染物）的腐烂真菌。其基因组有10个染色体，DNA长度为30Mb。

> *P. chrysosporium* 的基因组测序项目是由美国能源部承担的（http://www.jgi.doe.gov/programs/whiterot.htm）。我们在第5章中介绍了白腐病的BLAST服务器。
> 更多关于裂殖酵母基因组序列分析的详细信息可以在Wellcome Trust Sanger Institute的网站（http://www.sanger.ac.uk/Projects/S-pombe/）上得到。TIGR上面也提供了*Schizosaccharomyces*基因的索引（http://www.tigr.org/tdb/spgi/）。

15.4.6 裂殖酵母

裂殖酵母（*Schizosaccharomyces pombe*）的基因组大小为13.8Mb。其全基因组序列首先由一个欧洲联合会报道（Wood等，2002），基因组分为3个染色体（表15-9）。

表15-9 裂殖酵母基因组特征

染色体数目	长度/Mb	基因数目	平均基因长度/bp	编码率/%
1	5.599	2255	1446	58.6
2	4.398	1790	1411	57.5
3	2.466	884	1407	54.5
总基因组	12.462	4929	1426	57.5

注：来源于Wood等（2002）。

> Leland Hartwell、Timothy Hunt和Sir Purse Nurse因为在研究细胞周期控制方面的工作，一起分享了2001年的诺贝尔生理和医学奖。Nurse研究了裂殖酵母，Hartwell研究了啤酒酵母，而Hunt则研究了海胆和其他生物。参见http://www.nobel.se/medicine/laureates/2001/。

值得注意的是，它有4940个预测的蛋白质编码基因（包括11个线粒体基因）和33个假基因。实际上这比啤酒酵母基因的1/4还少，甚至是所有观察到真核生物中蛋白质编码基因数目最少的。事实上，一些细菌基因组编码的蛋白质都比它多，如*Mesorhizobium loti*（6752个预测基因）和*Streptomyces coelicolor*（7825个预测基因）。

裂殖酵母中的基因密度大约为每个基因2400个碱基对，比啤酒酵母的稍低。基因间的区域比较长，预测大约有4730个内含子。在啤酒酵母中，只有4%的基因有内含子。

裂殖酵母和啤酒酵母是在3.3亿～4.2亿年前分开的。一些基因和蛋白质序列在这两种真菌中被相等地分开，它们在真菌和人类中的同源蛋白也如此。你能够在NCBI基因组主页上用TaxPlot这个工具来直观地看到这一点（图15-23）。这样，就很容易发现真菌

基因和人类基因相当近源。通过比较分析能够阐明这些真菌生物的不同的基因基础，如裂殖酵母双裂变生殖时的倾向性和裂变酵母中转座子比较少。

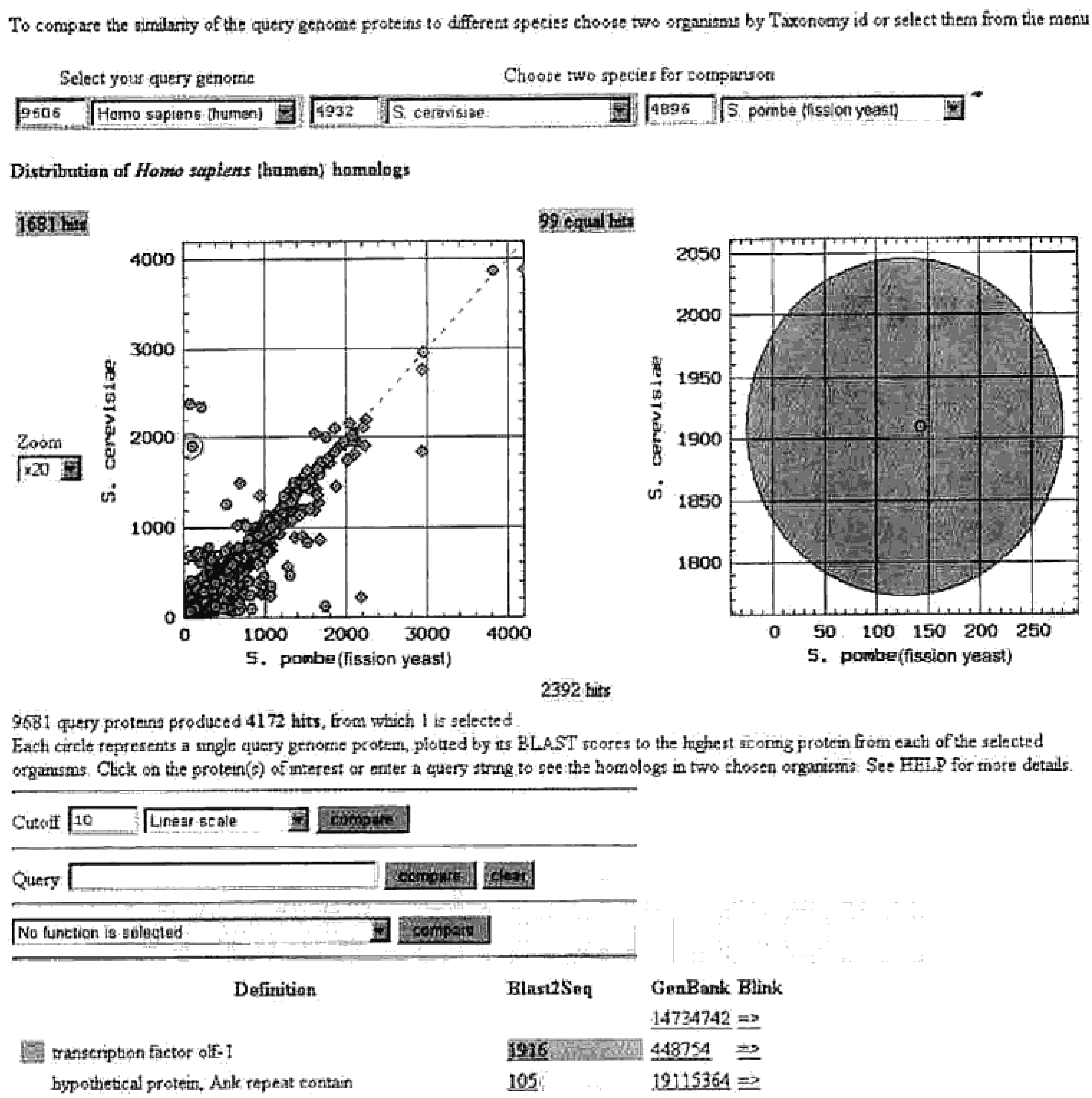


图 15-23　NCBI 上的 TaxPlot 工具表明许多啤酒酵母蛋白质和裂殖酵母蛋白质与人类蛋白质非常近源，但前两者却彼此不近源。能够用这个工具来比较不同物种的蛋白质组。选择一个物种作为参考（这里选择了人类），然后选择一个物种进行比较，这里我们从上面的下拉菜单中选择了真菌啤酒酵母和裂殖酵母。主要的输出结果是每个人类蛋白质和各个真菌蛋白质的 blastp 打分图。99 个匹配分值差不多，1682 个匹配和啤酒酵母更近源，2391 个匹配和分质酵母更近源。很容易得到和人类同源蛋白近源的真菌蛋白质的信息，例如，在图的左边点击一个数据点，我们就能看到这个蛋白质（在图的右侧和底部）是一个转录因子 olf-1（Wang 和 Reed，1993），人类蛋白和啤酒酵母的同源比对分值为 1916，但是人类蛋白和裂殖酵母的同源比对分值却只有 105

15.5　展望

出芽酵母 *S. cerevisiae* 是生物中最重要的模式生物之一，主要有如下原因。

① 它是第一个完成测序的真核生物基因组，并且它有相当紧密的基因组大小和结构。

② 作为一个单细胞真核生物，比起人类和其他多细胞生物它更为简单。

③ 生物研究者们已经掌握了酵母基因组学的很多知识，也收集了许多能够有效研究酵母基因功能的分子工具。基于对基因功能的全基因组范围分析的功能基因组学方法现在已经被采用了。例如，每一个基因（这种基因大于6000个）都被敲除并标签上了分子条形码，从而可以进行大规模并行的基因功能研究。

许多其他真菌基因组现在也在被测序。在生物的每一个分支，我们发现比较基因组学分析在很多方面都有着重要作用：寻找蛋白质编码基因（通过同源查找），如基因组复制分析等的进化研究，帮助我们揭示调节细胞生长的生化途径。

15.6 常见问题

对啤酒酵母这种重要的模式生物，我们必须认识到我们忽略的地方。单基因敲除的基因型是怎样导致其特殊的表现型呢？我们急于想知道人类基因突变是怎样导致疾病发生的，但是即使在像酵母这样所谓的简单模式生物中，我们都没有完全了解所有的蛋白质相互作用，而这正是细胞功能的基础。如果我们把基因组认为是一台机器的蓝图，现在我们已经有了部分“机器组件”——基因产物。接下来我们必须将这些部分拼在一起使得这台机器能够在不同背景下发挥作用。在一些像SGD和MIPS的酵母数据库中，有包括功能基因组学的研究结果在内的基因注释，这给功能分析提供了一个很好的起点。

15.7 网络资源

相关的网络资源见表15-10。

表15-10 关于一些真菌的网络资源

介绍	网址
酵母基因组测序项目	http://www.sanger.ac.uk/Projects/S_pombe/
	http://aspergillus-genomics.org/
	http://sequence-www.stanford.edu/group/candida/index.html
白色念珠菌信息	http://alces.med.umn.edu/Candida.html
真菌基因组资源	http://gene.genetics.uga.edu/
酵母启动子数据库	http://cgsigma.cshl.org/jian/
调控序列分析工具	http://www.ucmb.ulb.ac.be/bioinformatics/rsa-tools/
SAGE高级查询	http://genome-www.stanford.edu/Saccharomyces/SAGE/AdvancedQuery.html
酵母SAGE	http://www.sagenet.org/yeast/yeastintro.htm
UROFAN项目	http://www.mips.biochem.mpg.de/proj/eurofan/index.html
耶鲁基因组分析中心	http://ygac.med.yale.edu/
细胞周期分析项目	http://genome-www.stanford.edu/cellcycle/
	http://cmgm.stanford.edu/pbrown/explore/index.html
基因组范围的表达页面	http://web.wi.mit.edu/young/expression/
孢子形成时的转录项目	http://cmgm.stanford.edu/pbrown/sporulation/
真菌遗传学中心	http://www.fgsc.net/
酵母虚拟图书馆	http://genome-www.stanford.edu/Saccharomyces/VL-yeast.html
	http://www.bio.net:80/hypermail/YEAST/
蛋白质序列信息中心	http://mips.gsf.de/
MIPS酵母基因组	http://www.mips.biochem.mpg.de/proj/yeast/
	http://www.mips.biochem.mpg.de/cgi-bin/proj/expression/start.pl
酵母基因组数据库	http://genome-www.stanford.edu/Saccharomyces/
有关线粒体起源的数据库	http://www.proteome.com/databases/

15.8 问题讨论

1. 出芽酵母 *Saccharomyces cerevisiae* 有时候被描述为简单生物，因为它是单细胞的，其基因组编码一组相对较少的基因（大约6000个），它作为一种模式生物用来进行基因组学研究，但我们只了解了其一半基因的功能。现在有很多功能基因组学工具能够拿来应用，像酵母基因敲除库（如每个基因的无效等位基因）。你怎样利用这些功能基因组学工具来扩展我们对酵母基因功能的理解？

2. 真菌是多细胞生物（动物）的姐妹分支（图15-1）。你觉得通过研究真菌得到的有关基因组进化、基因功能和比较基因组学的一些规则能否在像人类、蠕虫和果蝇等多细胞生物上也适用？例如，我们讨论了一些真菌的全基因组复制，你是否认为人类基因组也进行同样的复制？在比较基因组学中，比起多细胞生物你是否觉得真菌之间的生物学特性更为相似？

15.9 问题

1. 酵母有多少个脂质运载蛋白？

2. ABC转运器是由一个大的跨膜蛋白家族构成的，这些蛋白水解ATP然后促使像氯化物等物质进行跨膜运输。酵母中有多少ABC转运器？

3. 使用啤酒酵母基因组数据库：

① 先访问SGD网站（http://genome-www4.stanford.edu/cgi-bin/SGD/seqTools）。

② 选一个没有鉴定的开放阅读框。用基因/序列资源（一个分析工具）来查找，选一个染色体（如十二号染色体），然后选择染色体特征表。列出的第一个假定的开放阅读框是YLL067C。

③ 尽量找一下这个开放阅读框的可能功能是什么。对于一些没有被鉴定的开放阅读框，只有相对很少的可用信息；但对于其他的能找到很多。在染色体特征表中选择“信息”（info）来查看与图15-12相似的网页。

—这个蛋白质的物理特征是什么（如分子大小、等电点）？

—这个蛋白质是否有已知结构域？

—是否已经发现这个蛋白和其他蛋白有无相互作用？

—在不同的生理状态下（如压力应答、孢子形成等），这个基因是被诱导的还是被抑制的？

④ 试着使用功能链接（在开放阅读框的信息网页的最下面），它会同时搜索6个数据库：

—Yeast Path Calling（双杂交分析）；

—SGD SAGE；

—Worm-Yeast protein comparison；

—Yeast Microarray Global Viewer；

—Yeast Protein Function Assignment；

—Triples。

⑤ 这个基因在哪些其他生物中也存在？比较一下在SGD中搜索和自己进行BLAST搜

索得到的结果，看哪个更好？

15.10 自测题

1. 你在 SGD（酵母基因组数据库）中鉴定出了一个酵母蛋白质，它来自开放阅读框 YDR477W。下面哪一个是你接下来能够找到有关这个蛋白质更多信息的最好的地方？

(a) MIPS；

(b) COG（Clusters of Orthologous Groups）；

(c) GenBank；

(d) KEGG（Kyoto Encyclopedia of genes and Genomes）

2. 啤酒酵母有很多理由成为一个诱人的模式生物。下面哪个不是酵母的一个有用的特征？

(a) 基因组比较小；

(b) 可以通过同源重组进行基因敲除；

(c) 大重复 DNA 序列片段能作为一个更高级的真核生物的模型；

(d) 高密度开放阅读框

3. 酵母基因组数据库不能实现下面中的哪一项？

(a) 显示许多酵母蛋白质中可能的蛋白质相互作用；

(b) 显示染色体和所有它们假定的开放阅读框；

(c) 提供和酵母蛋白质同源的细菌蛋白质列表；

(d) 提供酵母基因的基因芯片表达数据

4. 啤酒酵母基因组很小（编码大概 6000 个基因）。一般认为在大约 1 亿年前

(a) 整个基因组复制了，然后四倍化；

(b) 基因组经过了许多片段的复制，然后部分基因丢失；

(c) 整个基因组复制，然后基因丢失；

(d) 整个基因组复制，然后基因转变

5. 基因复制之后，复制基因对最为常见的命运是

(a) 两个拷贝都在基因组中间，仍然保持着相似的功能；

(b) 其中一个拷贝被删除；

(c) 一个拷贝经过突变然后变成一个假基因；

(d) 其中一个或者两个基因都在功能上分化了

6. 现在对于“功能基因组学”有很多定义，在下面选择最合适的一个

(a) 主要是根据基因芯片或者 SAGE 技术得到的基因组范围的基因表达数据鉴定基因的功能；

(b) 主要是根据对蛋白质相互作用和蛋白质网络的研究鉴定基因的功能；

(c) 结合了遗传学、生物化学和细胞生物学方法来研究一个基因的功能、其 mRNA 产物和它相应的蛋白质产物；

(d) 用全基因组范围的筛选和分析来鉴定基因的功能

7. 用转座子基因印迹（由 Patrick Brown 率先提出）方法主要的优点是

(a) 这个方法技术上很容易实现，并且能够被放大来研究多个基因的功能；

(b) 插入等位基因和敲除等位基因都能被研究；

(c) 用这种方法能够研究任何感兴趣的已知基因；

(d) 突变株能够被存放起来供其他研究者接下来的研究

8. “YKO”项目试图系统地敲除所有的酵母开放阅读框。这个方法一个潜在的缺陷是

(a) 对于酵母基因来说分子条形码可能有时候是有毒的；

(b) 这个方法不适合用来找新基因，主

要针对的是已知基因；

(c) 基因敲除突变株不能保存用来给其他科研工作者研究；

(d) 突变可能不是无效的

9. 裂殖酵母基因组最重要的特征之一是

(a) 据预测它编码的蛋白质少于 5000 个，这使得它的基因组（或者蛋白质组）甚至比一些细菌基因组都小；

(b) 预测的内含子数目和预测的开放阅读框的数目差不多；

(c) 它和啤酒酵母有同源基因，它也有许多细菌基因的同源基因；

(d) 它的基因组大小和啤酒酵母差不多，尽管这两个物种在几百万年前就分开了

10. 酵母是否为由美国食品和药物管理局（FDA）批准的惟一的主要研究生物

(a) 是；

(b) 不是

15.11 推荐读物

Guarro 和 Stchigel（1999）发表了一个对于真菌分类非常精彩的综述。这篇综述还从临床角度介绍了真菌。Baldauf 等（2000）对真核生物进行了更为详尽的分类，这将会在第 16 章具体讨论（图 16-1）。在此，我们期望用部分系统发生分析（图 15-1）来介绍真菌及它们和动物之间的亲密关系。

另外一篇有关啤酒酵母基因组的优秀综述是由 Johnston（2000）撰写的。这篇文章讨论了一些鉴定酵母基因功能的方法。

参 考 文 献

Aalto, M. K., Ronne, H., and Keranen, S. Yeast syntaxins Sso1p and Sso2p belong to a family of related membrane proteins that function in vesicular transport. *EMBO J.* **12**, 4095-4104 (1993).

Ainsworth, G. C. Fungus infections (mycoses). In K. F. Kiple, ed. *The Cambridge World History of Human Disease*. Cambridge University Press, New York, 1993, pp. 730-736.

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., and Doolittle, W. F. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290**, 972-977 (2000).

Braun, E. L., Halpern, A. L., Nelson, M. A., and Natvig, D. O. Large-scale comparison of fungal sequence information: Mechanisms of innovation in *Neurospora crassa* and gene loss in *Saccharomyces cerevisiae*. *Genome Res.* **10**, 416-430 (2000).

Bulloch, W. *The History of Bacteriology*. Oxford University Press, New York, 1938.

Casselton, L., and Zolan, M. The art and design of genetic screens: Filamentous fungi. *Nat. Rev. Genet.* **3**, 683-697 (2002).

Cherry, J. M., et al. Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature* **387**, 67-73 (1997).

Chu, S., et al. The transcriptional program of sporulation in budding yeast. *Science* **282**, 699-705 (1998).

Dacks, J. B., and Doolittle, W. F. Novel syntaxin gene sequences from *Giardia*, *Trypanosoma* and algae: implications for the ancient evolution of the eukaryotic endomembrane system. *J. Cell Sci.* **115**, 1635-1642 (2002).

DeRisi, J. L., Iyer, V. R., and Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-686 (1997).

DeRisi, J., et al. Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants. *FEBS Lett* **470**, 156-160 (2000).

Dujon, B., et al. Complete DNA sequence of yeast chromosome XI. *Nature* **369**, 371-378 (1994).

Dwight, S. S., et al. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.* **30**, 69-72 (2002).

Garbarino, J. E., and Gibbons, I. R. Expression and genomic analysis of midasin, a novel and highly conserved AAA protein distantly related to dynein. *BMC Genomics* **3**, 18 (2002).

Giaever, G., et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387-391 (2002).

Goffeau, A., et al. Life with 6000 genes. *Science* **274**, 546, 563-567 (1996).

Guarro, J., Gene J., and Stchigel, A. M. Developments in fungal taxonomy. *Clin. Microbiol. Rev.* **12**, 454-500 (1999).

Harrison, P. M., Kumar, A., Lang, N., Snyder, M., and Gerstein, M. A question of size: The eukaryotic proteome and the problems in defining it. *Nucleic Acids Res.* **30**, 1083-1090 (2002).

Holstege, F. C., et al. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717-728 (1998).

Jelinsky, S. A., and Samson, L. D. Global response of *Saccharomyces cerevisiae* to an alkylating agent. *Proc. Natl. Acad. Sci. USA* **96**, 1486-1491 (1999).

Johnston, M. The yeast genome: On the road to the Golden Age. *Curr. Opin. Genet. Dev.* **10**, 617-623 (2000).

Johnston, M., et al. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XII. *Nature* **387**, 87–90 (1997).

Katinka, M. D., et al. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414**, 450–453 (2001).

Kinsey, J. A., Garrett-Engele, P. W., Cambareri, E. B., and Selker, E. U. The *Neurospora* transposon Tad is sensitive to repeat-induced point mutation (RIP). *Genetics* **138**, 657–664 (1994).

Krumlauf, R., and Marzluf, G. A. Genome organization and characterization of the repetitive and inverted repeat DNA sequences in *Neurospora crassa*. *J. Biol. Chem.* **255**, 1138–1145 (1980).

Kuchenmeister, F. *On animal and vegetable parasites of the human body, a manual of their natural history, diagnosis, and treatment.* Sydenham Society, London, 1857.

Kumar, A., et al. An integrated approach for finding overlooked genes in yeast. *Nat. Biotechnol.* **20**, 58–63 (2002a).

Kumar, A., et al. The TRIPLES database: A community resource for yeast molecular biology. *Nucleic Acids Res.* **30**, 73–75 (2002b).

Lowe, T. M., and Eddy, S. R. A computational screen for methylation guide snoRNAs in yeast. *Science* **283**, 1168–1171 (1999).

Lynch, M., and Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).

Mackiewicz, P., et al. How many protein-coding genes are there in the *Saccharomyces cerevisiae* genome? *Yeast* **19**, 619–629 (2002).

Marc, P., Devaux, F., and Jacq, C. yMGV: A database for visualization and data mining of published genome-wide yeast expression data. *Nucleic Acids Res.* **29**, E63–3 (2001).

Margulis, L., and Schwartz, K. V. *Five Kingdoms. An Illustrated Guide to the Phyla of Life on Earth.* W. H. Freeman and Company, New York, 1998.

Mewes, H. W., et al. Overview of the yeast genome. *Nature* **387**, 7–65 (1997).

Mewes, H. W., et al. MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34 (2002).

Ohno, S. *Evolution by Gene Duplication*. Springer Verlag, Berlin, 1970.

Oshiro, G., et al. Parallel identification of new genes in *Saccharomyces cerevisiae*. *Genome Res.* **12**, 1210–1220 (2002).

Perkins, D. D., Radford, A., and Sachs, M. S. *The Neurospora Compendium: Chromosomal loci.* Academic Press, San Diego, CA, 2001.

Piskur, J. Origin of the duplicated regions in the yeast genomes. *Trends Genet.* **17**, 302–303 (2001).

Protopopov, V., Govindan, B., Novick, P., and Gerst, J. E. Homologs of the synaptobrevin/VAMP family of synaptic vesicle proteins function on the late secretory pathway in *S. cerevisiae*. *Cell* **74**, 855–861 (1993).

Ross-Macdonald, P., et al. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413–418 (1999).

Roth, J. F. The yeast Ty virus-like particles. *Yeast* **16**, 785–795 (2000).

Rozen, S., et al. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**, 873–876 (2003).

Sankoff, D. Gene and genome duplication. *Curr. Opin. Genet. Dev.* **11**, 681–684 (2001).

Selker, E. U. Premeiotic instability of repeated sequences in *Neurospora crassa*. *Annu. Rev. Genet.* **24**, 579–613 (1990).

Seoighe, C., and Wolfe, K. H. Updated map of duplicated regions in the yeast genome. *Gene* **238**, 253–261 (1999).

Smith, V., Botstein, D., and Brown, P. O. Genetic footprinting: A genomic strategy for determining a gene's function given its sequence. *Proc. Natl. Acad. Sci. USA* **92**, 6479–6483 (1995).

Smith, V., Chou, K. N., Lashkari, D., Botstein, D., and Brown, P. O. Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* **274**, 2069–2074 (1996).

Spellman, P. T., et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.* **9**, 3273–3297 (1998).

Wagner, A. Robustness against mutations in genetic networks of yeast. *Nat. Genet.* **24**, 355–361 (2000).

Wagner, A. Birth and death of duplicated genes in completely sequenced eukaryotes. *Trends Genet.* **17**, 237–239 (2001).

Wang, D. Y., Kumar, S., and Hedges, S. B. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B Biol. Sci.* **266**, 163–171 (1999).

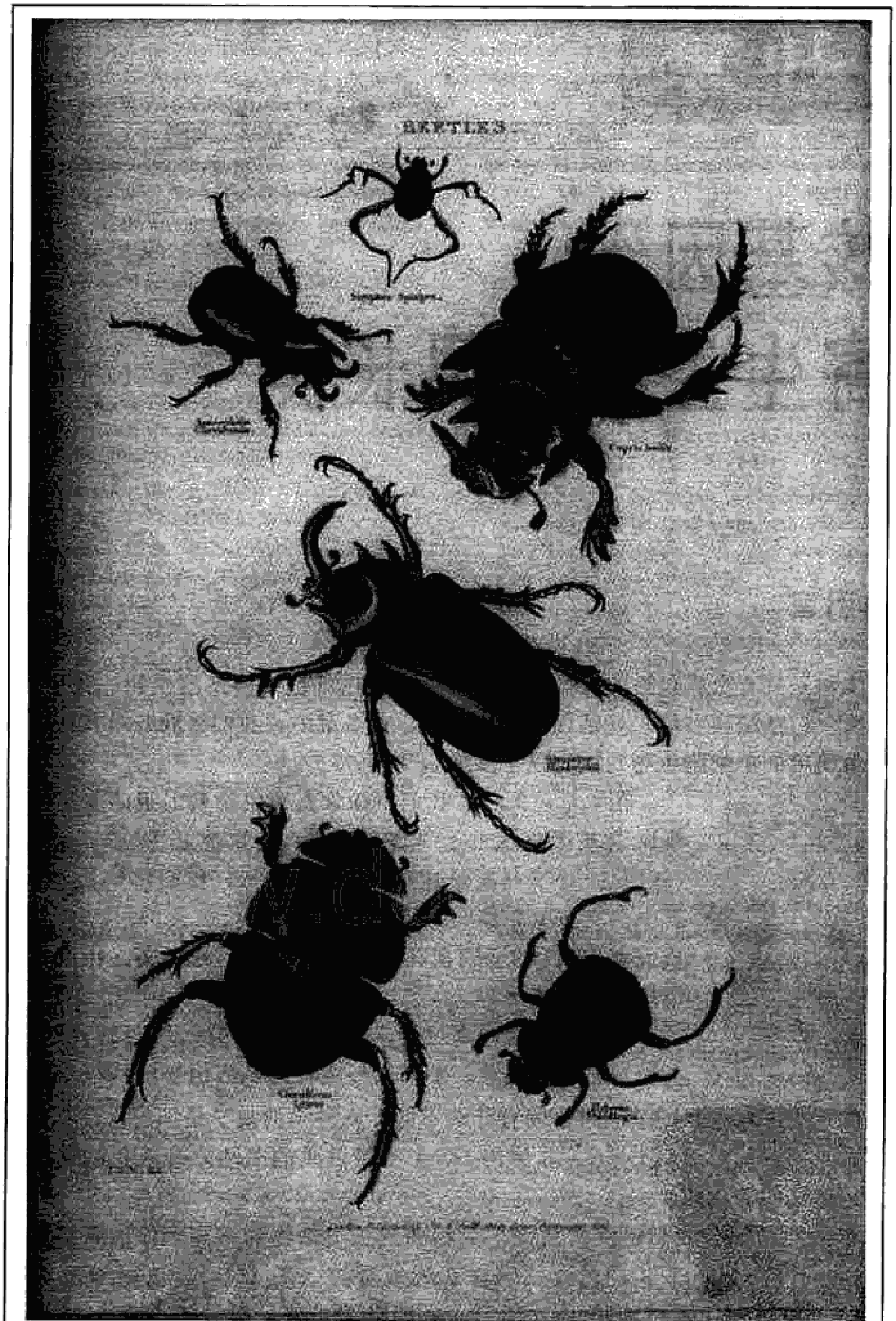
Wang, M. M., and Reed, R. R. Molecular cloning of the olfactory neuronal transcription factor Olf-1 by genetic selection in yeast. *Nature* **364**, 121–126 (1993).

Winzeler, E. A., et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).

Wolfe, K. H., and Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997).

Wong, S., Butler, G., and Wolfe, K. H. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc. Natl. Acad. Sci. USA* **99**, 9272–9277 (2002).

Wood, V., et al. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871–880 (2002).



在第 16 章我们将从生物信息学角度来讨论多种多样的真核生物。Baron Georges Cuvier（1769～1832）试图对动物进行系统的分类。他将动物界分成了 4 个大类：脊椎动物、软体动物、节肢动物和枝条动物（也叫植物性动物）。这幅图画描绘了几种甲虫（Cuvier，1849，图版 XXXI）。从 18 世纪以来，有 350000 种甲虫被详细描述，但是据估计大约存在着数以百万计的甲虫种类

第16章

真核基因组：从寄生生物到灵长类

16.1 引言

真核生物是具有由膜包被的核结构和细胞骨架的单细胞或多细胞生物。我们在第15章介绍的真菌（包括酿酒酵母）就是属于真核生物。在本章中，我们将展开对真核生物的广泛讨论，从最简单的单细胞生物到植物以至多细胞动物。

eukaryotes的同义词包括了eucaryotae、eucarya、eukarya和eukaryotae，这个词来源于希腊词eu（真实的）和karutos（有核，指的是细胞核）。

在本章前部分，我们将探索真核生物基因组的普遍特性。我们将讨论基因组大小、非编码DNA（如重复DNA）和编码DNA（基因）。对于一段给定的基因组DNA，我们将解决注释的问题：其中含有多少非编码的DNA？分别是什么类型？有多少编码蛋白质或者RNA的基因在其中？什么是基因的定义？什么软件程序可以用来探测基因？

更广泛地说，通过对基因组DNA的探索，我们可以回答生物和物种进化的问题。不同真核生物的基因组是如何以染色体的形式组织起来的？小鼠和大鼠的基因组之间有什么不同？新种类的蚊子是如何从已有的蚊子种类中产生的？除了我们对编码DNA、非编码DNA和蛋白质的关注之外，我们还将研究真核生物的染色体结构和功能。

在本章的后半部分我们将探索从寄生生物到灵长类动物的不同个体的真核基因组。我们引用Baldauf等（2000）构建的真核生物系统发育树（图16-1）。该树通过对4个蛋白的序列进行分析得到：延长因子1α（EF-1α）、肌动蛋白、α-微管蛋白和β-微管蛋白。我们已经在第15章中讨论了真菌的基因组，它们属于与多细胞生物（动物）很相近的一个大组。我们将从下向上讨论该树中具有代表性的生物。它们包括二倍单胞体生物贾第虫和包括疟原虫在内的其他原生生物，还包括第一个被全基因组测序的拟南芥和水稻在内的植物，以及多细胞生物（从蠕虫到昆虫，再到鱼类和哺乳动物类）。我们将在第17章讨论人类基因组。

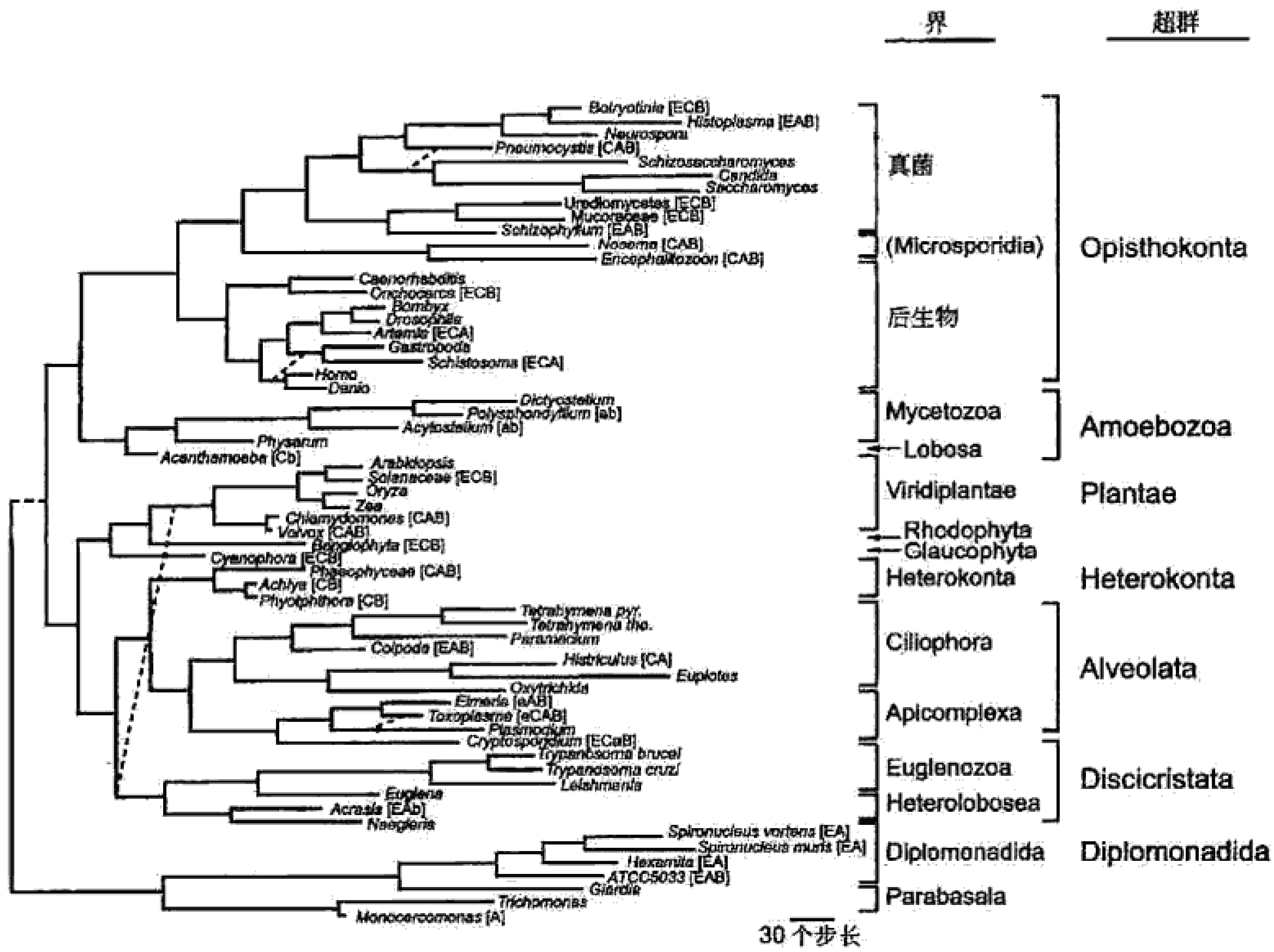


图 16-1 基于蛋白质序列分析得到的真核生物进化树。被用于构建进化树的蛋白是 EF-1α（在图中用 E 表示）、肌动蛋白（C）、α-微管蛋白（A）和 β-微管蛋白（B）。该树可以同使用核糖体 RNA 小亚基序列构建的生命树（图 12-1）的真核部分进行比较。在这棵树中，14 个生物界被分成了 7 个大的组。其中的一个组（Opisthokonta）包括真菌、小孢子虫（第 15 章）和多细胞动物（脊椎动物和无脊椎动物）。本树是通过对密码子第二个位置上的核苷酸进行最大简约度和最大似然度分析得到的。对于数据缺失的生物类别，序列用括号标出［例如（EAB）］来自 Baldauf 等（2000），有修改，已获授权

本章的讨论焦点之一是真核生物的全基因组测序。为什么我们需要获得全基因组的序列呢（线虫测序组织，1998）？

① 全基因组序列为发现基因组内编码的所有基因提供了基础。其他的方法（如 EST 法）不够全面和广泛。

② 比较基因序列可以发现基因的结构和调控元件。

③ 它为研究一个物种的分子进化以及物种内个体、种群，以及物种和物种之间的差异程度提供了必要的基础。

④ 它为将来的实验提供了一系列的工具。

在对真核生物进行全基因组测序的同时，另一个分子生物方法是对来自成百上千不同种类生物的某些特定基因进行鉴定。模式基因的使用是对模式生物的使用的补充。例如，在 GenBank 中搜索核酮糖-1,5-双磷酸羧化酶（rubisco，*rbcL*）可以得到超过 15000 条结果。*rubisco* 是植物的一个重要基因（将在后面讨论）。从各种物种中得到的分子序列数据的具备为系统发育树的构建和结构研究提供了必要的条件。其他被广泛研究的基因包括了高度保守的分子，如肌动蛋白、微管蛋白和其他一些蛋白（表 11-2）。

16.2 真核生物的普遍特性

16.2.1 真核生物与原核生物的主要不同

真核生物和原核生物有着共同的祖先，但是当我们比较它们时可以发现许多显著的不同之处（Cavalier-Smith，2002；Vellai 和 Vida，1999；Watt 和 Dean，2000）。一些不同之处在表 16-1 中被标出。

有性生殖被称为配子配合，其过程是雄性和雌性配子的单倍染色体组合形成杂合子（如受精卵）。

① 虽然真核生物和原核生物的生命形式都有很大的多样性，但是很少有细菌或者古细菌是可以被人的肉眼所看见的。许多真核生物是单细胞的微生物，但是相对于原核生物，我们通常会把真核生物同多细胞生物（如植物和动物）联系在一起。

② 真核生物具有原核生物所不具备的三个特性：a. 一个由膜包被的核体；b. 一个由细胞内膜包被形成的广泛存在的细胞器系统；c. 细胞骨架，包括了肌动蛋白、微管蛋白和分子马达等元件。另外，原核生物缺少产生能量的细胞器，不能进行细胞内吞。通过内吞作用，细胞外的物质可以进入细胞内（Vellai 和 Vida，1999）。

③ 大部分真核生物进行有性繁殖，虽然有一些进行无性繁殖。细菌缺少接合体的融合，不通过性交换 DNA。

④ 真核基因组的大小分布很广泛，跨越了 5 个不同的数量级（表 16-2）。相反，大多数古细菌和细菌的基因组都处于 0.5～13Mb（见第 12 章和第 14 章）。

⑤ 原核生物基因组中编码蛋白质的基因的密度很大，而非编码 DNA 和重复序列很少。例如，大肠杆菌基因组的 0.7%由非编码重复序列组成（Blattner 等，1997）。相反，真核基因组的大部分都是由非编码 DNA 组成。表 16-1 给出了一些例子。

表 16-1 细菌和真核基因组特征

特征	*E. coli*K-12	Parasite①	Yeast②	Slime Mold③	Plant④	Human⑤
基因组大小/Mb	4.64	22.8	12.5	8.1	115	3289
GC 比例/%	50.8	19.4	38.3	22.2	34.9	41
基因数目	4288	5268	5770	2799	25498	30000～40000
基因密度/kb·基因$^{-1}$	0.95	4.34	2.09	2.60	4.53	27
编码比例	87.8	52.6	70.5	56.3	28.8	1.3
内含子个数	0	7406	272	3578	107784	53295

①*Plasmodium falciparum*；②*Saccharomyces cerevisiae*；③*Dictyostelium dicoideum*；④*Arabidopsis thaliana*；⑤*Homo sapiens*。

注：1. kb 表示千碱基对；Mb 表示兆碱基对。

2. 数据来源于 Gardner 等（2002），Blattner 等（1997），国际人类基因组测序联盟（2001）。

⑥ 原核生物是单倍体的。真核生物可以是单倍体、双倍体或者其他倍数性。这种高倍数性质给予了真核生物丰富的进化机制，如杂合优势（Watt 和 Dean，2000）。

⑦ 基因组的组织方式不同。真核生物的基因组被组织成线形的染色体（而大部分细菌和古细菌的基因组是环形的，见图 14-1）。这些真核染色体有一个着丝点，在两端各有一个端粒。这些特性在原核生物的染色体中没有，虽然类似着丝粒的元件已经被报道（Ben-Yehuda 等，2002；Moller-Jense 等，2002）。细菌分离 DNA 的机制尚未明了。

表 16-2 真核生物基因组大小

分类	门类	基因组大小/Gb	基因组比例
所有真核细胞	—	0.003～686	228667
圆藻	—	—	22333
	Apicomplexians	0.009～201	22333
	纤毛虫	0.024～8.62	359
	腰鞭毛	1.37～98	72
硅藻		0.035～24.5	700
阿米巴		0.035～686	19600
眼虫		0.098～2.35	24
真菌/微粒子虫		0.003～1.47	490
动物	—	—	3325
	海绵	0.059～1.78	30
	丝孢子虫	0.227～1.83	8
	昆虫	0.089～9.47	106
	软骨鱼	1.47～15.8	11
	多骨鱼	0.345～133	386
	两栖动物	0.93～84.3	91
	爬行动物	1.23～5.34	4
	鸟类	1.67～2.25	1
	哺乳动物	1.7～6.7	4
	Placozoa	0.04	
植物	—	—	6140
	绿藻	0.080～30	375
	蕨类植物	0.098～307	3133
	裸子植物	4.12～76.9	19
	被子植物	0.050～125	2500

注：1. 数据来自 Graur 和 Li（2000）动物基因组大小数据库（T. R. Gregory）（http://www.genomsize.com）和国家生物技术信息中心（http://www.ncbi.nlm.nih.gov）。

2. 0.001Gb 等于 1Mb。用皮克值乘以 0.9869×10^{3} 以得到 Gb 数值。

16.2.2 *C* 值矛盾：为什么真核基因组大小差异如此之大

在真核生物基因组中，单倍体基因组大小（*C* 值）有很大的差异。表 16-2 显示了不同类别的真核生物的 *C* 值，表 16-3 给出了某些特定的真核生物的 *C* 值。有些基因组很小，例如小孢子虫 *Encephalitozoon cuniculi*（2.9Mb，第 15 章）。有一些生物的基因组大致为几十亿个碱基对。*C* 值的差异也在单细胞生物中存在，如阿米巴原虫高达 20000 倍之多。在动物界中，这个差异可以达到约 3000 倍。

值得注意的是，*C* 值的范围与生物的复杂程度并没有相关性。有些生物［如拟南芥（一种植物）和 *Fugu rubripes*（一种河豚）］具有很紧凑的基因组，但是它们的近亲生物在具有相似的复杂度的同时却有着大好几个数量级的基因组。这种相关性缺失的现象叫做 *C* 值矛盾（Hartl，2000；Knight，2002；Hancock，2002；Kidwell，2002）。近年来多种真核生物的基因组序列以草稿或最终版本的形式被测出，包括秀丽隐杆线虫（1998）、黑腹果蝇（2000）、人类（2001）以及小鼠（2002）（见下文以及第 12 章）。这些被测序的全基因组序列给 *C* 值矛盾提供了一个明白的答案：基因组内充满了非编码 DNA 序列。这导致了基因组大小的不同。我们将在以后讨论非编码 DNA 序列。

表 16-3 真核生物 C 值表

种类	通用名	C值/Gb
Saccharomyces cerevisiae	酵母	0.012
Neurospora crassa	河豚	0.043
Dysidea crawshagi	海绵	0.054
Caenorhabditis elegans	线虫	0.097
Drosophila melanogaster	果蝇	0.12
Paramecium aurelia	纤毛虫	0.19
Oryza sativa	水稻	0.47
Strongylocentrotus purpuratus	海胆	0.80
Gallus domesticus	鸡	1.23
Erysiphe cichoracearum	霉菌	1.5
Boa constrictor	蛇	2.1
Parascaris equorum	蛔虫	2.5
Carcharias obscurus	鲨鱼	2.7
Canis familiaris	狗	2.9
Rattus norvegicus	大鼠	2.9
Xenopus laevis	非洲蟾	3.1
Homo sapiens	人	3.3
Nicotania tabacum	烟草	3.8
Locusta migratoria	飞蝗	6.6
Paramecium caudatum	纤毛虫	8.6
Allium cepa	洋葱	15
Truturus cristatus	蝾螈	19
Thuja occidentalis	西洋松	19
Coscinodisucus asteromphalus	硅藻	25
Lilium formosanum	百合	36
Amphiuma means	蜥蜴	84
Pinus resinosa	红松	68
Protopterus aethiopicus	肺鱼	140
Amoeba proteus	阿米巴虫	290
Amoeba dubia	阿米巴虫	690

注：来源于 Graur 和 Li（2000）；NCBI（http://www.ncbi.nlm.nih.gov）；Cameron 等（2000）；以及基因组大小数据库（http://www.cbs.dtu.dk/databases/DOGS/abbr.table.common.txt）。

16.2.3 许多真核基因组包含非编码和重复的 DNA 序列

细菌和古细菌的基因组内包括了基因和基因间的小片段。通常原核基因组是环形的，每1000个碱基对的长度中含有一个基因（第14章和表16-1）。相反地，真核生物基因组包含

C 值通常用碱基对数目或者皮克（pg）来度量。1pg DNA 相当于 1Gb 个碱基对。

真核生物中，正常生殖细胞是单倍体而体细胞是二倍体。因此，同一个生物体内的不同细胞可以具有不同的倍数性。倍数性是指一个细胞内染色体的套数。

倍数性可以有许多不同的表现方式，有些单细胞真核生物（如酿酒酵母）可以在单倍体或者二倍体状态下生长。三倍体的果蝇可以存活但是生殖能力降低。虽然我们把体细胞和生殖细胞的倍数性分开对待，但是一些体细胞同样可以有多种倍数性。例如，在人体内的一小部分肝细胞通常是三倍体。通常地，在哺乳动物中，额外多出的一条染色体也会导致机体的死亡。

一个植物 C 值的在线数据库（由 M.D Bennett 和 I.J. Leitchj 建立）在 http://www.rbgkew.org.uk/cval/homepage.html。动物基因组大小数据库（来自 T. Ryan Gregory）在 http://www.genomesize.com/。另一个有关基因组大小的数据库（DOGS）为 http://www.cbs.dtu.dk/database/DOGS/。

了小部分的蛋白质编码基因和大量的非编码 DNA 序列。这些非编码序列包括了重复 DNA、编码具有调控功能的 RNA 的基因、外显子之间的将被切除出成熟 mRNA 的内含子。

对 C 值矛盾的其他解释并不合理。真核生物基因组中编码蛋白质的基因在 10 倍数量级上下变动，但是这个差异要小于基因组的大小差异。同样，不同物种间 mRNA 长度的差异也不能解释 C 值矛盾，因为基因的平均长度和基因的大小之间没有相关性。

真核生物基因组中的非编码序列的很大一部分由重复性 DNA 序列组成。许多不同长度的重复的核苷酸序列遍布整个基因组（Jurka，1998）。我们也将在人类基因组中讨论这些重复性序列（第 17 章）。在哺乳动物中，多达 60％的基因组 DNA 是重复的，而在酵母中也达到了 20％。

Britten 和 Kohne（1968）做了一些早期的实验，揭示了真核基因组中的重复 DNA 序列。他们从很多物种中提取基因组 DNA，剪切并把双链分离。在一定的温度和盐度下，经过一定的时间，DNA 重新复性。他们发现不少真核基因组 DNA 重新组合成许多分离的小片段，而细菌或者病毒的基因组没有这种情况发生。大量的真核生物 DNA 异常迅速地重新聚合。对于小鼠的基因组来说，基因组 DNA 的 10％快速聚合，包含了大约 100 万份拷贝（图 16-2，箭头 A）。这些高度重复性的 DNA 序列分布在染色体的密集部分，被称为异染色质（Redi 等，2001；Avaramova，2002）。另外的 20％重新组成了包括 1000～100000

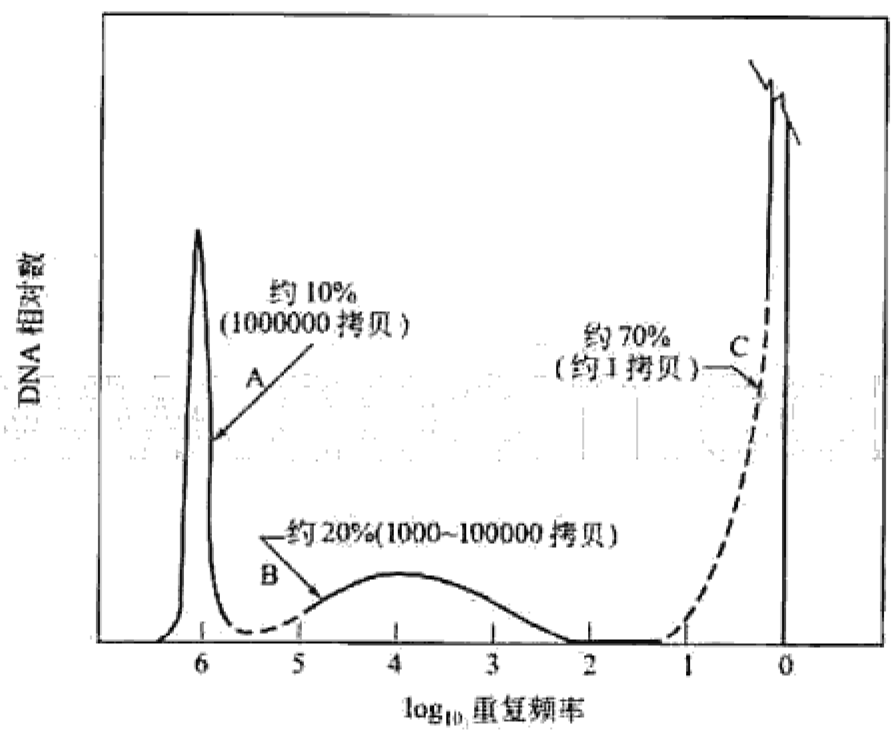


图 16-2 基因组 DNA 的复杂度可以用 DNA 的变性和复性实验来估计。该图（Britten 和 Kohne，1968）描述了小鼠基因组 DNA 的相对数量（*y* 轴），横轴为 DNA 重复频率的常用对数。这些数据是从 C_0t 曲线推出，描述了在特定时间内特定 DNA 浓度下重组的基因组 DNA 的百分含量。如果 $C_0t_{1/2}$ 越大，则重组反应越慢。三种类型已被发现：速度最快的组分由 10％的小鼠基因组 DNA 组成（箭头 A），它表示了高度重复的卫星序列 DNA；速度处于中等的组分由 20％的基因组 DNA 组成，包含了拷贝数从 1000～100000 的重复序列；最慢的 70％基因组 DNA 是单拷贝的 DNA。Britten 和 Kohne（1968）也在其他真核生物的实验中取得了类似结果，虽然不同物种间有一些明显不同。图表授权引用

Britten和Kohne于1968年使用了多种不同的技术将单链DNA从双链DNA中解离出来，例如羟磷灰石色谱（一种磷酸钙柱），将带有放射性标记的DNA结合到滤层上固定DNA，以及分光光度学。DNA重新组合的速度是培养时间（t）和DNA浓度（C_0）的函数。C_0t曲线表明了仍然为单链形式存在的DNA的比例，这是图16-2中展示的数据的基础。

反转录转座子（也叫反转座子或反转座元件）是一种可以换位的元件，它通过一个反转录的过程，并且利用RNA中间体，将自己拷贝到基因组的多个位置上，这个过程与反转录病毒相似。

份不同DNA片段的一组（箭头B）。最后，70%的DNA序列是惟一的，只含有一个拷贝（箭头C）。这些DNA组成常染色质，在染色体里处于并不高度密集的部位，可以被转录。染色体的带型（图16-19）对应着异染色质和常染色质的区域。

研究重复性序列的起源和功能是非常有趣的。它们有哪些不同类型？它们是从何时何地起源的？它们的重复是有一定逻辑的还是它们只是无目的地复制自身？我们还只是刚刚开始了解重复性DNA的性质，包括人类基因组。重复性DNA有时候也被称为“垃圾DNA”或者“自私的DNA”，因为它们不断向整个基因组扩张。但是，重复性DNA可能在染色体结构、重组事件和某些基因的功能中起重要作用（Makalowski，2000；见下文）。

真核基因组中的重复性DNA主要分为5种（IHGSC，2001；Jurka，1998；Kidwell，2002；Makalowski，2000）。

① 散布的重复性DNA（转座子导致的重复）。散布的重复DNA组成了人类基因组的大约45%（见第17章）。这些重复DNA可以由拷贝RNA中间产物（反转录元件）或者DNA中间产物（DNA转座子）生成（表16-4）。当mRNA被反转录后插入到基因组中，基因也可以被拷贝到反转录转座的位置。这些基因可以被鉴定出来，因为它们缺少内含子，同时它们两侧含有小片段的重复的序列。表16-5给出了一些哺乳动物中被反转录转座的基因的例子。

散布性的重复序列可以分成4个类（Ostertag和Kazazian，2001；Kidwell，2002）（见图17-18）。

a. 长末端重复（*LTR*）转座子，是由RNA反转录而成的元件。它们也成为类反转录病毒元件。*LTR*转座子在两端有长达数百个碱基对的*LTR*。

表16-4 转座子元件种类

类别	子类	超家族	家族举例	大致范围/bp
反转录转座子	LTR转座	Ty1-copia	Opie-1(maize)	3000～12000
	非LTR转座	LINEs	*LINE-1*(human)	1000～7000
		SINEs	*Alu*(human)	100～500
DNA转座子	剪-贴转座	Mariner-Tc1	*Tc1* in *C. elegans*	1000～2000
		P	*P* in *Drosophila*	500～4600
	圆型转座	*Helitrons*	*Helitrons* in *A. thaliana*, *O. sativa*, and *C. elegans*	5500～17500

注：数据来自Kidwell（2002），已获许可。

Barbara McClintock被授予了1983年的诺贝尔奖，因为她在玉米（*Zea mays*）中发现了可以移动的基因元件。你可以在http://www.nobel.se/medicine/laureates/1983/找到关于这个开创性工作的更多阅读材料。Wojciech Makalowski提供了一个网站，其中描述了许多SINEs和LINEs的例子，同时提供了关于它们的作用的假说（http://www.ncbi.nlm.nig.gov/Makalow/sines.html）。

用“retropseudogene”对Entrez进行核酸序列搜索产生65条记录（2003年1月）；用“retrotransposed”产生10条记录，但是用“retrotransposon”搜索可以产生11000条记录。

表 16-5 由转座产生的基因举例

转座基因			原基因			分布	年龄/MYA
名称	RefSeq	Chr	名称	RefSeq	Chr		
ADAM20	NM_003814	14q	*ADAM9*	NM_003816	8p	Human, not macaque	<20
Cetn1	NM_004066	18p	*Cetn2*	NM_004344	Xq28	Mammals	>75
Glud2	NM_012084	Xq	*Glud1*	NM_005271	10q	Human, not mouse	<70
Pdha2	NM_005390	4q	*Pdha1*	NM_000284	Xp	Placentals	约 70
SRP46	NM_032102	11q	*PR264/SC35*	NM_003016	17q	Human, simians	约 89
Supt4h2	NM_011509	10	*Supt4h*	NM_009296	11	Mouse	<70

注：1. 转座基因缺少内含子，两侧序列重复。

2. Chr 表示染色体；MYA 表示百万年；*ADAM* 表示一种蛋白酶和金属酶；*Cetn*（cent）表示 EF-手蛋白；*Glud* 表示谷氨酸脱氢酶；*Pdha2* 表示 α_2-丙酮酸脱氢酶；*Supt4h* 表示 Ty4 抑制子（酵母）。

3. 来源于 Betran 和 Long（2002）（见有关基因的论文和其他文献）及在 LousLink 和 Entrez（NCBI）用"retropseudogene"搜索得到的结果。

b. 长散布元件（*LINEs*），编码了具有反转录活性的酶，还可能编码了其他蛋白质。在哺乳动物中，*LINE1* 和 *LINE2* 是最常见的。

c. 短散布元件（*SINEs*），也是由 RNA 产生的元件。在灵长类动物中发现的 *Alu* 重复序列，这是 *SINEs* 的著名例子。我们将在图 16-8 中见到 *Alu* 重复序列的一个例子。

d. DNA 转座子，占人类基因组的大约 3%。

② 被修饰的假基因。这些基因并不被转录或者翻译（Harrison 和 Gerstein，2002；Echols 等，2002）。它们曾经是有功能的基因，但是它们缺乏蛋白质产物。因为它们有一个终止密码子或者由阅读框移位导致的中断，所以它们可以被识别出来。假基因可以由反转录转座事件产生（如由有反转录酶活性的 *LINEs* 导致的随机插入事件），或者由基因的复制产生然后失去其基因功能。

③ 简单序列重复。这些微卫星序列（通常长度为 1～6bp）和小卫星序列（通常长度为 12～500bp 的重复）包含了短的序列，例如（A)*n*、(CA)*n* 或者（CGG)*n*（我们将在图 16-8 中见到人类基因组 DNA 的一个 CA 重复序列的例子）。复制过程中的滑移是产生简单重复序列的一个可能原因。

简单重复序列在不同的物种中具有不同的长度和位置偏好性。例如，(AT)*n* 在拟南芥中很普遍，(CT/GA)*n* 经常在线虫中出现（Schlotterer 和 Harr，2000）。在人类基因组中，简单重复序列引起了人们的极大兴趣，因为它们通常在不同个体间体现了多态性，可以用来作为遗传学标记。另外，一些三碱基的重复序列（如 CAG）与 14 种疾病相关，包括亨廷顿症（Huntingon's disease）（Cummings 和 Zoghbi，2000）。我们将在第 18 章中讨论这个问题。

④ 片段复制。这些复制的片段长度为 1～200kb 或 300kb，被从基因组的一个位置拷贝到另一个位置（Bailey 等，2001），出现在同一个染色体上或者不同染色体之间。Bailey 等估计出人类基因组包含了大约 3.6%的复制区域（长度至少 1kb，并具有 90%～98%的序列相似性）。

我们将考虑位于人类染色体 9 上的脂质运载蛋白的一个基因群作为片段性复制的一个例子。所有生物的脂质运载蛋白被分成了 14 个单元的分支（Gutierrez 等，2000）。人类基因组中至少含有 10 个脂质运载蛋白基因，分布在染色体 9 的 q32～34 位置上。图 16-3 给出了一个基因组 DNA 的图示，包括了眼泪脂质运载蛋白（LCN1）和气味分子结合蛋白的基因（修改自 Lacazette 等，2000）。基于他们对该基因组区域的分析，Lacazette 等（2000）提出了一个解释脂质运载蛋白基因和在染色体 9 p34 观察到的假基因的模型（图 16-3 底部）。假设的脂质运载蛋白的共同祖先有 7 个外显子和 6 个内含子（图 16-3 顶部），这是哺乳动物脂

质运载蛋白的典型结构（Salier，2000）。这个基因发生了串联复制（图16-3，步骤1）。然后两个古代基因进行了功能上的分化（步骤2）。该基因座发生了2次复制，产生了LCN1和OBPⅡ两个旁系同源蛋白（步骤3）。然而，现在的该基因座上只有两个*OBP*Ⅱ基因（图16-3底部），*LCN1*基因伴随着两个假基因（*LCN1b*和*LCN1c*）。由此推断可能发生了部分

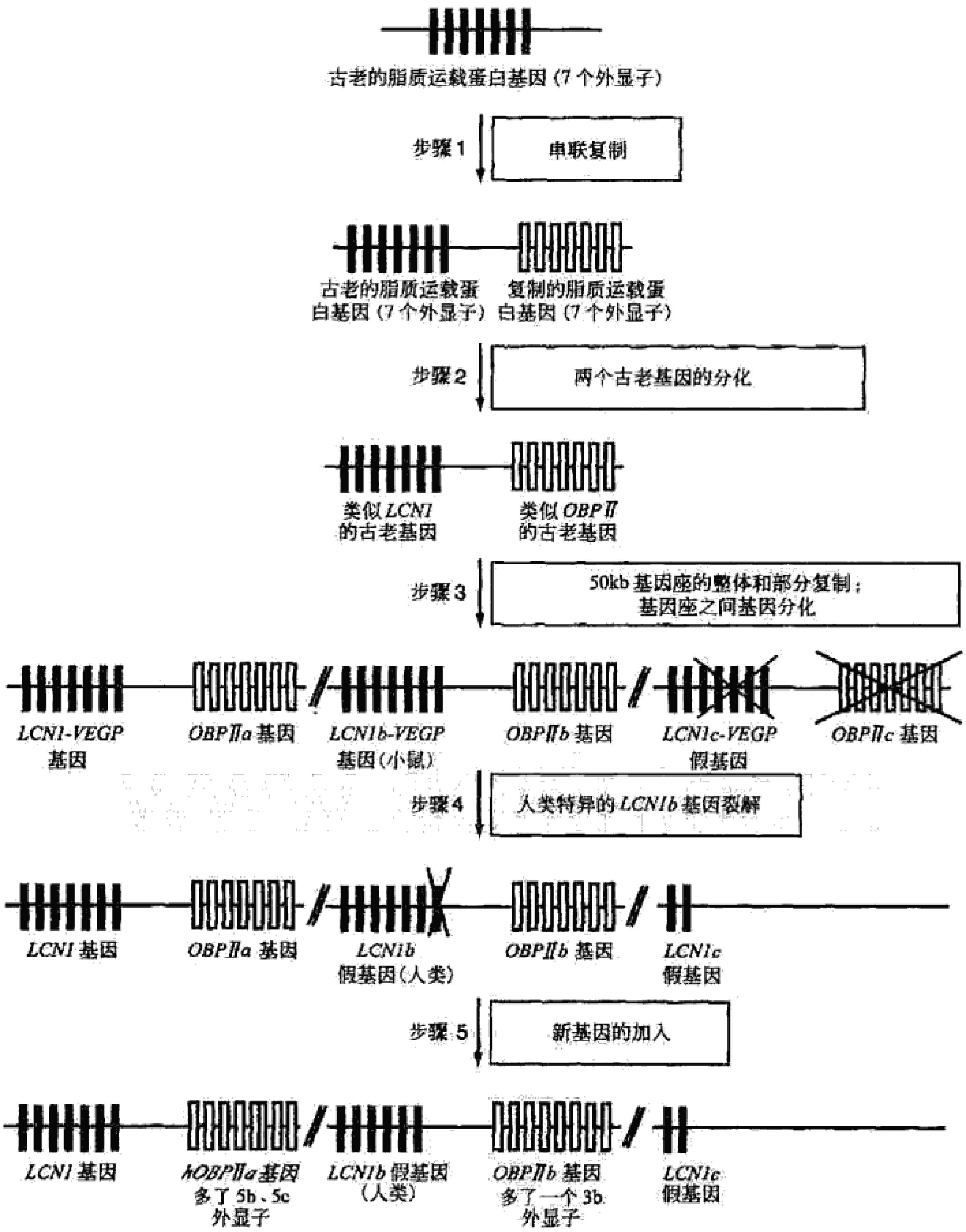


图16-3 串联复制引起的基因进化。Lacazette等在2000年提出这个模型来解释一个假想的载脂蛋白基因是如何进化而来的，并且这个模型能解释现在的基因模式和通过基因座分析所确定的假基因。首先，一个古老的脂质运载蛋白的7个外显子进行了复制（步骤1），然后进行了功能上的分化（步骤2）。这个区域包含了2个基因，发生了2次复制（步骤3），接着一个基因被删除（假想的*OBP*Ⅱ*c*基因），*LCN1*的部分基因被删除（步骤4）。最后，多个新的外显子加入（步骤5）。已获授权使用本图

复制（图 16-3，步骤 3），随之发生的是人类基因组中 *LCN1b* 的复制，但是在小鼠中并没有发生复制（步骤 4）。最后，人类基因组中的两个新外显子——*OBP* Ⅱ *a* 和 *OBP* Ⅱ *b* 表明了不同的外显子的复制是有选择性的（步骤 5）。

Mark Gerstein 的实验室提供了一个关于假基因的网站（http://www.pseudogene.org）。它包括人类、蠕虫、果蝇、酵母和植物中的假基因的列表和对它们的描述。
小鼠基因组包含了一个编码甘油醛-3-磷酸脱氢酶的基因 *Gapdh*，它在 19 个染色体上至少有 400 个假基因（小鼠基因组测序组织，2002）。到目前（2003 年 1 月），*Gaphd* 基因被认为存在于小鼠的第 7（小鼠基因组测序组织）、第 6（NCBI 的 LocusLink 和 UniGene）和第 5 个染色体上。由于大量假基因的存在导致了精确定位的困难。
一些作者认为微卫星序列的长度为 1～6bp 而另一些认为是 1～12bp。
如果读者想看简单重复序列的例子，请去 Entrez 的核酸库搜索 “microsatellite”，将有多达 50000 条记录供挑选。
Tandem Repeats Finder 是一个在线工具，它可以在一个序列内进行长达 2000bp 的串联重复序列的搜索（http://tandem.biomath.mssm.edu/trf/trf.html）（Benson，1999）。
Case Western Reserve 大学（CWRU）复制浏览器可以让你在人类基因组中鉴定片段复制。读者可以从 http://humanparalogy.gene.cwru.edu/SDD/获取。它利用加州大学 Santa Cruz 分校的基因组浏览器（我们将在第 17 章讨论）。CWRU 的这个网址提供了多达 8500 多个人类基因组中的片段复制。

⑤ 串联重复序列块，例如在端粒、着丝粒和核糖体基因群附近的串联重复序列块。端粒是在真核生物染色体末端发现的重复性序列。它们维持染色体的稳定性，防止被降解，同时阻止染色体末端之间的融合。表 16-6 列出了多个端粒的重复序列。在人类的端粒中，TTAGGG 序列被重复了几千次。如果试用 TTAGGG TTAGGG TTAGGG TTAGGG 进行 blastn 的搜索，限制搜索结果为人类的基因组，去除低复杂度的过滤设置，得到的结果将包含几千个记录，其中大部分都来自于端粒，类似于图 16-4 中所述。

表 16-6 真核生物端粒重复序列

种　类	端粒重复	参考文献
Arabidopsis thaliana，其他植物	TTTAGGG	McKnight 等，1997
Ascaris suum（线虫）	TTAGGC	Jentsch 等，2002
Euplotes aediculatus，*Euplotes crassus*，*Oxytricha nova*（纤毛虫）	TTTTGGGG	Jarstfer 和 Cech，2002；Shippen-Lentz 和 Blackburn，1989；Melek 等，1994
Giardia duodenalis，*Giardia lamblia*	TAGGG	Upcroft 等，1997；Hou 等，1995
Guillardia theta（隐滴虫）	$[AG]_7AAG_6A$	Douglas 等，2001
Homo sapiens，其他脊椎动物	TTAGGG	Nanda 等，2002
Hymenoptera，*Formicidae*（蚂蚁）	TTAGG	Lorite 等，2002
Paramecium，*Tetrahymena*	TTGGGG，TTTGGG	McCormick-Graham 和 Romero，1996
Plasmodium falciparum	AACCCTA	Gardner 等，2002
Plasmodium yoelii yoelii	AACCCTG	Carlton 等，2002

着丝点是染色体上一个重要的位置，它为纺锤体的微管丝提供了附着点，使得在细胞进行有丝分裂和减数分裂时染色体能分离开来（Choo，2001）。所有的真核生物染色体都有有功能的着丝点，虽然其序列可能因物种的不同而不同。在人类基因组中，着丝点的 DNA 包含了长为 171bp 的 α 卫星 DNA 序列重复，共 1～4Mb。几乎所有的真核生物着丝点都可以与组蛋白 H3 相关的蛋白（在脊椎动物中称为 CENP-A）结合。这个蛋白质-DNA 复合体在染色质的着丝点位置形成了一种特定结构，这种结构对着丝点（纺锤体纤维结合的位置）的功能行使是必不可少的。

```
>gi|7407196|gb|AF236885.1|AF236885 Homo sapiens clone p10 chromosome 6
telomeric repeat region
GGATCCCCCCCAACTCATGACTGTCGGGCTATTTCCAGGCCGCATCGACAGTGAACAAAATCCTTTCTGT
TTGCAGCCCTGAATAATCAGGGTTAGGGTTAGGGTTAGGGGTTAGGGGTTGGGGTTGGGGTTAGGGTTAG
GGTTGGGTTAGGGTTAGGGTTAGGGGTTAGGGTTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTAGGGTC
AGGGTCAGGGTCAGGGTCAGGGTCAGGGTTAAGGGTCAGGGTCAGGGTCAGGGTCAGGGTCAGGGGTTAG
GGTTAGGGTTAGGGTCAGGGTCAGGGTCAGGGTCAGGGTCAGGGGTTAGGGTCGGGGTCGGGGTCGGGGT
CAGGGGTCAAGGGTCAAGGGTCGGGGTCAGGGGTCAAGGGTCGGGGTCGGGGTCCGGGTCAGGGTGAGGG
TGAGGGTGAGGGTGAGGGTGGGGG
```

图 16-4 真核生物端粒重复序列的一个例子。对人类 DNA 用 (TTAGGG)，进行 blastn 搜索得到许多结果，包括了图示的这条（GenBank 编号 AF236885）。TTAGGG 重复序列是红色（图中为黑色）和被下划线标出。其他的相关的 TCGAAA 被下划的虚线标出。成千上万的这种重复序列占据了端粒

位于人类染色体 9 p32～34 的脂质运载蛋白包括了 α-1-微球蛋白/Bikunin（NM_001633）、补足物组分 8，γ 肽链（NM_000606）、脂质运载蛋白 1（比小牛血清蛋白迁移速度快的泪前清蛋白，LCN1）、脂质运载蛋白 2（癌基因 *24p3*）（NM_005564）、气味分子结合蛋白（OBP）2A（NM_014582）和气味分子结合蛋白 2B（NM_014581）、血清类黏蛋白 1（NM_000607）和血清类黏蛋白 2（NM_002571），以及前列腺素 D2 合成酶（NM_000954）。见 Chan 等（1994）和 Dewald 等（1996）。

LCN1 的功能还不清楚，它是从小鼠的泪腺克隆出的 cDNA。小鼠和小牛的 OBPs 能选择性结合多种化学分子（如萜烯、乙醛、酯类和麝香）（Pevsner 等，1990；Pelosi，1996）。通常人们推断人类的 *OBP* 基因产物也可以转运疏水性小分子。

小鼠基因组测序组织（2002）描述了在小鼠 X 染色体上有 8 个脂质运载蛋白基因而灵长类动物基因组中没有，这可能是通过局部基因的复制导致的。

端粒重复序列是由端粒酶合成的。端粒酶是一种核蛋白，具有反转录活性。一个关于端粒的数据库（TelDB）位于 http:///www.genlink.wustl.edu/teldb/index.html。

GenBank 中关于人类 α 卫星 DNA 序列的一个一致序列编号是 X07685。利用这个一致序列使用 blastn 进行搜索可以得到 20000 多条结果。如果你在结果中将人类排除在外（使用命令 "satellite NOT human [organism]"），你将发现人类的 α 卫星 DNA 序列也以与灵长类的序列匹配上。但是，人类 α 卫星 DNA 序列不是非灵长类动物的序列的同源物。

RepBase Update 是由 Jerzy Jurka 和他的同事们于 1990 年开发的。RepeatMasker 是由 Arian Smit 和 Phil Green 编写的。它们可以在以下地址获得：http://www.geospiza.com/products/tools/repeatmasker.htm。

基因信息研究所（GIRI）的 Censor 服务器在 http://www.girinst.org/。

我们使用的基因组 DNA 来自 contig NT.008769。

着丝粒附近的卫星序列是所有真核生物的共同特征，只有两个已知的例外。一个例外是，在酵母基因组中，整个着丝粒序列一共只有几百个碱基对的长度。第二个例外是新着丝粒——一种异常的着丝粒，它有功能正常的着丝点，在有丝分裂中稳定，但是缺少 α 卫星 DNA 序列（Amor 和 Choo，2002）。大约 60 例人类新着丝粒被发现，大部分有三染色体性或四染色体性（多余的染色体拷贝）。

16.2.4 检测重复性 DNA 的软件：RepeatMasker 和 Censor

在真核基因组 DNA 中寻找重复序列（如 SINEs 和 LINEs），对于基因组分析来说是非常必要的。知道重复性序列的位置可以帮助鉴定非编码区域，而重复序列自身也是"分子化石"，可以用在比较基因组学研究中（第 17 章）。寻找重复性序列的一个实用方法是对数据库进行搜索，寻找那些已知的重复序列和低复杂度区域。RepBase Update 是真核基因组重

复序列的首要数据库。多个软件程序（如 RepeatMasker 和 GIRI）的 Censor Server 都可以针对提交的序列对数据库进行搜索（Smit，1999；Jurka，2000）。

为了鉴定和屏蔽重复性序列，可以使用 RepeatMasker 网站服务器。表 16-7 列出了多个服务器的地址，你也可以在本地的机器上安装数据库和程序。我们将用包含 *RBP4* 基因在内的人类 10 号染色体上的 100000bp 的基因组 DNA 序列来讨论这个问题。把你的序列粘贴到一个提供了输出选项的文本框内（图 16-5）。GIRI 的 censor 服务器同样提供了类似的界面（图 16-6）。

表 16-7 提供基因组 DNA 重复序列鉴定的软件的网络资源

程　序	描　述	网　址
RepeatFinder	一个分析基因组序列重复结构的计算分析系统	http://www.tigr.org/softlab/
RepeatMasker	University of Washington Genome Center	http://repeatmasker.genome.washington.edu
RepeatMasker	位于 EMBL 的服务器	http://woody.embl-heidelberg.de/repeatmask/
RepeatMasker	位于 Barcelona 的服务器	http://humangen.med.ub.es/tools/RepeatMasker.html
RepeatMasker	用于斑马鱼，位于 the Wellcome Trust Sanger Instiutte	http://www.sanger.ac.uk/Projects/D_rerio/fishmask.shtml
Censor Server	Genetic Information Research Institute	http://www.girinst.org/

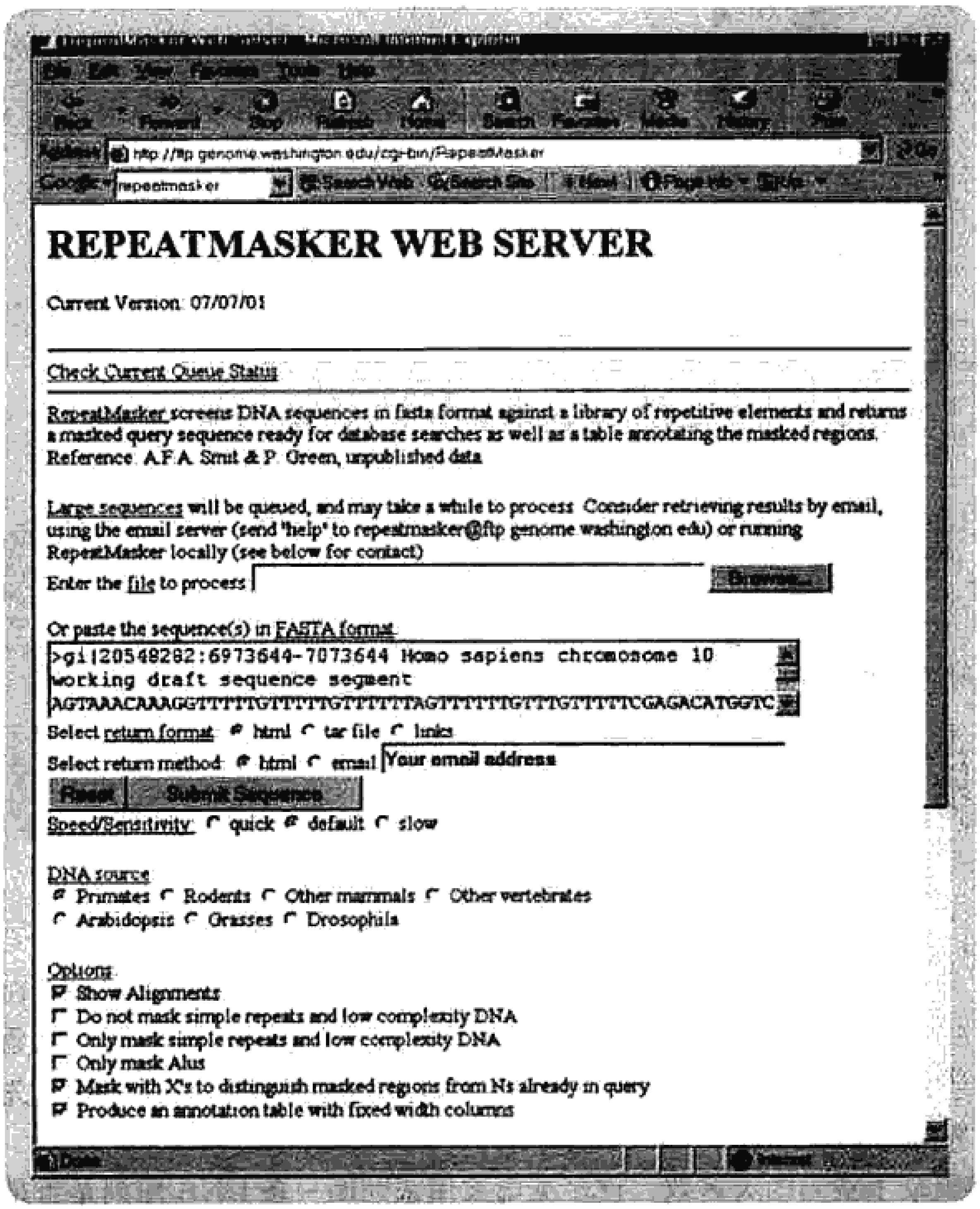


图 16-5 RepeatMasker 服务器可以用来鉴定基因组 DNA 中的重复性序列，图中的输入序列是人类染色体 10 号上的 10000bp DNA 序列

BLAST 使用 SEG 和/或 DUST 程序来定义和屏蔽重复 DNA 序列，也能检测和屏蔽低复杂度的蛋白质序列（第 4 章）。

RepeatMasker 的输出信息包含了用 Smith-Waterman 算法得出的一个按照分值排列的列表、重复序列的位置和关于重复序列类型的信息（如 SINE/Alu、*LTR* 或者简单重复序列）（图 16-7）。两两的序列比对结果也被显示出来（图 16-8）。输入的序列也用 FASTA 的格式输出，其中的重复性序列用 X 或者 N 屏蔽（图 16-9）。这个版本的序列对于后续的数据库搜索尤其有用。最后，RepeatMasker 程序对找到的重复序列提供了一个总结（图 16-10）。

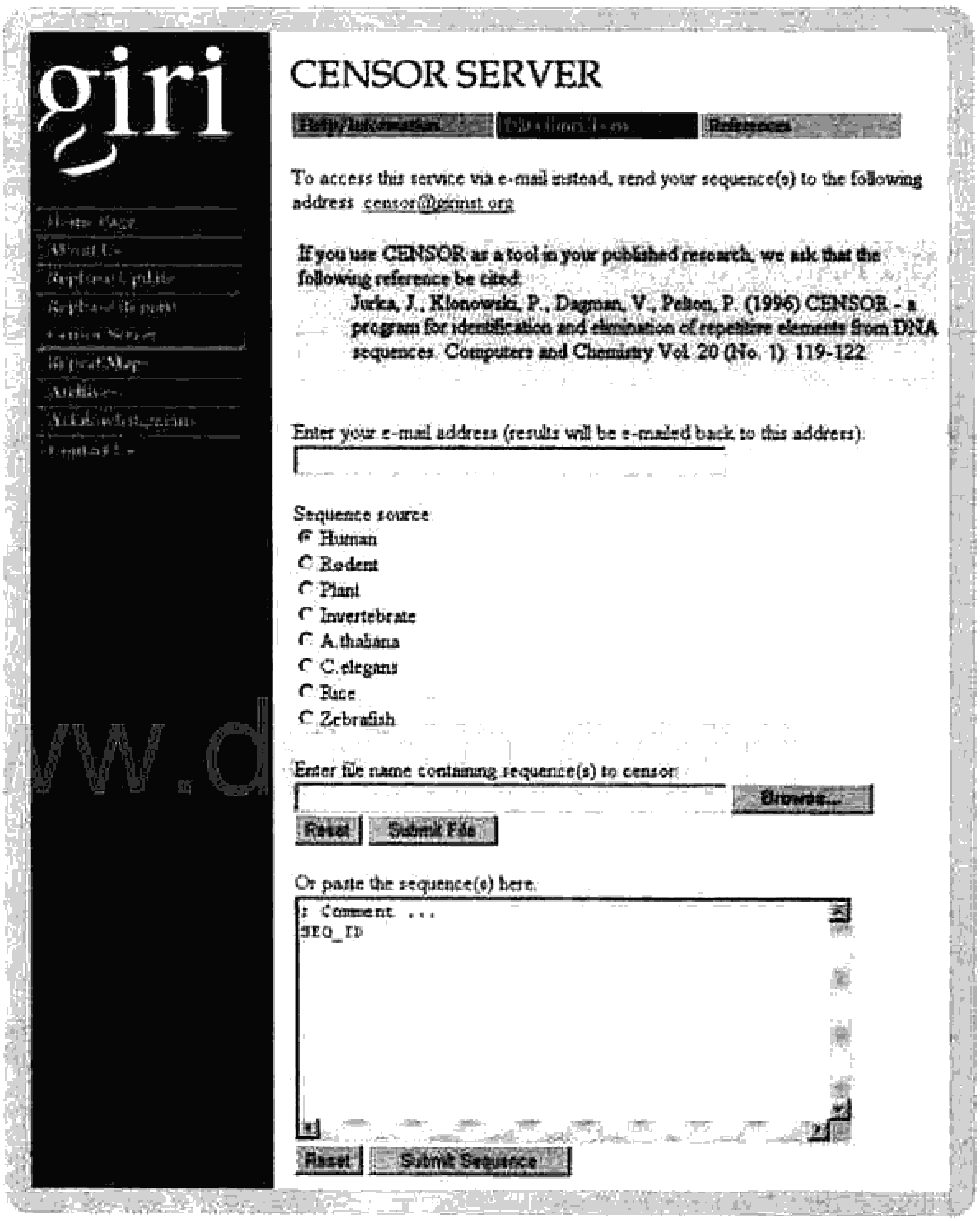


图 16-6 GIRI 的 censor 服务器，类似 RepeatMasker，可鉴定基因组 DNA 中的重复序列

16.2.5 基因的定义

我们从考虑非编码 DNA 和重复性序列开始探讨真核生物的基因组。基因组的编码蛋白质的部分更能引起人们的兴趣。因为它们决定了所有生物物种的大部分表型。理解真核生物基因组的两个最大挑战就是定义什么是基因和鉴定基因组 DNA 中包含的基因。我们将首先定义基因，然后给出鉴定基因的准则。

① 编码蛋白质的基因组成了基因的大多数。许多准则被用来判断一条 DNA 序列是否是

RepeatMasker completed 05-Nov-2002 12:15:50 PST
Repeat sequence:

SW score	perc div.	perc del.	perc ins.	position in query begin	end	(left)		matching repeat	repeat class/family	position in repeat begin	end	(left)	ID	
1446	13.8	2.8	10.4	19	223	(99287)	C	AluJo	SINE/Alu	(1)	311	137	1	
2438	7.3	0.3	0.3	224	525	(98985)	C	AluYa5	SINE/Alu	(9)	302	1	2	
1446	13.8	2.8	10.4	526	637	(98873)	C	AluJo	SINE/Alu	(175)	137	18	1	
823	14.8	0.0	2.3	1025	1152	(98358)	C	FLAM_C	SINE/Alu	(19)	126	1	3	
251	32.5	3.6	8.3	1201	1361	(98149)	+	MIR	SINE/MIR	9	173	(89)	4	
2180	13.5	0.7	0.0	1362	1665	(97845)	+	AluSq	SINE/Alu	1	306	(7)	6	
251	32.5	3.6	8.3	1666	1749	(97761)	+	MIR	SINE/MIR	173	259	(3)	4	*
684	30.3	5.2	0.9	1690	1920	(97590)	C	MIR	SINE/MIR	(16)	246	6	7	
392	22.2	0.0	1.3	2514	2612	(96898)	C	MLT1I	LTR/MaLR	(0)	450	319	8	
2335	10.1	0.0	2.8	2705	3022	(96488)	C	AluSq	SINE/Alu	(4)	309	1	10	
380	19.4	14.7	2.3	3033	3161	(96349)	C	MLT1J2	LTR/MaLR	(272)	178	34	11	
314	26.1	9.2	2.5	3354	3472	(96038)	+	MER34B	LTR/ERV1	5	131	(434)	12	
186	27.0	0.0	0.0	3474	3536	(95974)	+	(TGGG)n	Simple_repeat	2	64	(0)	13	*
588	24.4	0.0	0.0	3538	3709	(95801)	+	(TGGA)n	Simple_repeat	1	180	(0)	14	
215	26.5	0.0	0.0	3710	3758	(95752)	+	(TGGG)n	Simple_repeat	4	52	(0)	15	
363	20.2	5.0	0.5	3871	3956	(95554)	+	MER34C	LTR/ERV1	320	407	(168)	12	
2026	14.8	2.1	0.0	3957	4246	(95264)	C	AluJb	SINE/Alu	(14)	298	3	17	
363	20.2	5.0	0.5	4247	4384	(95126)	+	MER34C	LTR/ERV1	407	544	(31)	12	
2161	10.3	1.0	0.3	4896	5186	(94324)	+	AluSp	SINE/Alu	1	293	(20)	20	
337	10.6	0.0	0.0	5355	5428	(94082)	C	Alu	SINE/Alu	(0)	296	223	22	
248	6.8	11.4	0.0	5423	5466	(94044)	+	MADE1	DNA/Mariner	31	79	(1)	23	*
386	24.1	7.5	1.5	5474	5606	(93904)	C	MLT1F	LTR/MaLR	(0)	542	402	24	
231	16.7	0.0	6.2	5624	5671	(93839)	C	MLT1F2	LTR/MaLR	(389)	206	162	24	
2134	9.7	0.0	4.4	5674	6002	(93508)	C	AluSp	SINE/Alu	(0)	316	1	27	
2046	10.7	0.0	0.0	6003	6272	(93238)	C	AluSq	SINE/Alu	(13)	300	31	28	
320	29.3	4.1	1.6	6281	6403	(93107)	C	MLT1F2	LTR/MaLR	(436)	126	1	24	
221	36.2	9.6	0.0	6555	6692	(92818)	C	MIR	SINE/MIR	(66)	188	38	30	
233	21.9	12.5	0.0	6912	6975	(92535)	+	L1ME4a	LINE/L1	5530	5601	(520)	34	
213	21.1	0.0	0.0	7187	7224	(92286)	+	(CA)n	Simple_repeat	1	38	(0)	33	
459	25.1	7.2	6.0	7335	7566	(91944)	+	L1ME4a	LINE/L1	5791	6030	(91)	34	
2413	9.2	0.0	0.3	7567	7872	(91638)	C	AluSg	SINE/Alu	(5)	305	1	35	
459	25.1	7.2	6.0	7873	7958	(91552)	+	L1ME4a	LINE/L1	6030	6113	(8)	34	
215	29.7	1.8	4.4	8060	8240	(91270)	C	MIR	SINE/MIR	(27)	235	41	36	
443	26.9	4.9	5.8	8496	8718	(90792)	+	MLT1K	LTR/MaLR	310	530	(61)	36	

图 16-7 RepeatMasker 的输出表明了重复 DNA 序列（如 Alu、LINE 和 SINE）的不同得分和位置

一个编码蛋白质的基因。最原则性的要求是它必须含有一个长度至少为 90bp（对应 30 个编码氨基酸的密码子，或者大约分子量 3ku 的蛋白质）的开放阅读框（ORF）。

② 假基因（上文已经讨论过）并不编码有功能的基因产物。

③ 除了假基因外，还有其他许多种类的基因，它们也不编码蛋白质，而是编码许多有功能的 RNA 产物（Eddy，2001，2002）。这些基因包括转运 RNA（tRNA）基因。它们将信息从 mRNA 的三联密码子翻译成为氨基酸。有一个叫 tRNAscan-SE 的程序能鉴定基因组 DNA 中 99%～100%的 RNA 基因，而它的错误率是每 15Gb 产生 1 个假阳性（Low 和 Eddy，1997）。图 16-11 展示了 tRNAscan-SE 服务器的一个例子。

> tRNAscan-SE 搜索服务器位于 http://www.genetics.wustl.edu/eddy/tRNAscan-SE/。一个相关的资源是基因组 tRNA 数据库（GtRDB），它位于 http://rna.wustl.edu/GtRDB/。这个数据库总结了多种真核生物、细菌和古细菌的数量和种类。

④ 核糖体 RNA（rRNAs）基因也在翻译中执行重要功能。

⑤ 小核仁 RNA（snoRNAs）在核内执行功能。

⑥ 小核 RNA 在剪切体中起作用，它把初生的 RNA 中的内含子移除。

⑦ 微小 RNA（miRNAs）是大约 21～25 个碱基长度的 RNA，在各个物种中保守，可能作为其他 RNA 的反义调节者（Amros，2001；Ruvkun，2001）。

在基因组 DNA 的注释过程中，人们通常注重那些编码蛋白质的基因。但是，现在知道编码各种 RNA 的基因也有多种多样的重要功能。另外，鉴定非编码 RNAs 并不容易（Eddy，2002）。它们的全长也许很小（如微小 RNA），也没有 ORF 来帮助定义非编码基因的边界。非编码 RNA 的数据库搜索的覆盖率也许会比编码蛋白质的基因低，因为氨基酸的打分

```
186 26.98 0.00 0.00 gi|20548282:6973644-7073644 3474 3536 (95974)
C (CCCA)n#Simple_repeat (116) 64 2 * 5

  gi|20548282:6      3474 TGGATGTGTGGGTGAATGGGCAGCTGGATGGATGAGTGGGCAGGTAGATA 3523
                             i  v       ii    ii v   i   i  i     ii   i i i
C (CCCA)n#Simple_      64 TGGGTGGGTGGGTGGGTGGGTGGGTGGGTGGGTGGGTGGGTGGGTGGGTG 15

  gi|20548282:6      3524 AGTGGGTGGATGG 3536
                          i        i
C (CCCA)n#Simple_      14 GGTGGGTGGGTGG 2

Transitions / transversions = 7.50 (15 / 2)
Gap_init rate = 0.00 (0 / 63), avg. gap size = 0.00 (0 / 0)

    213 21.05 0.00 0.00 gi|20548282:6973644-7073644 7187 7224 (92286)
    (CA)n#Simple_repeat 1 38 (142) 5

      gi|20548282:6      7187 CACACACACACACTCATGCATGCACACACATATGCACA 7224
                                           v  ii  ii        i ii
      (CA)n#Simple_re       1 CACACACACACACACACACACACACACACACACACACA 38

    Transitions / transversions = 7.00 (7 / 1)
    Gap_init rate = 0.00 (0 / 38), avg. gap size = 0.00 (0 / 0)

2046 10.74 0.00 0.00 gi|20548282:6973644-7073644 6003 6272 (93238)
C AluSq#SINE/Alu (13) 300 31 1

  gi|20548282:6      6003 TTTATTTACTTATTTTTGAGACGGAGTTTCACTTTTGTTTCCCAGACTGG 6052
                             v   vi  v                  i  i     v     i
C AluSq#SINE/Alu      300 TTTTTTTTTTTTTTTTTGAGACGGAGTTTCGCTCTTGTTGCCCAGGCTGG 251

  gi|20548282:6      6053 AGTGCAATGGCGCCATCTTGGCTCAGTGCAACCTCTGCCTCCCAGGTTCA 6102
                                i      v    i      v         i       i
C AluSq#SINE/Alu      250 AGTGCAGTGGCGCGATCTCGGCTCACTGCAACCTCCGCCTCCCGGGTTCA 201

  gi|20548282:6      6103 AGCGATTCTCCTGCTTCAGCCTCCCGAGTAGCTGGGATTACAGGCGCGTG 6152
                                        i                                vi
C AluSq#SINE/Alu      200 AGCGATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGATTACAGGCGCCCG 151

  gi|20548282:6      6153 CCATCATGCCTGGCTAATTTTTGTATTTTTTGTAGAGACGGGGTTTCACC 6202
                             i  i   i                   v
C AluSq#SINE/Alu      150 CCACCACGCCCGGCTAATTTTTGTATTTTTAGTAGAGACGGGGTTTCACC 101

  gi|20548282:6      6203 ATATCGGCCAGGCTTATCTTGAACTACTGACCTGAGGTGATCCGCCCGCC 6252
                            i i         vi   i     v       v
C AluSq#SINE/Alu      100 ATGTTGGCCAGGCTGGTCTCGAACTCCTGACCTCAGGTGATCCGCCCGCC 51

  gi|20548282:6      6253 TCAGCCTCCCAAAGTGCTGG 6272
                            i
C AluSq#SINE/Alu       50 TCGGCCTCCCAAAGTGCTGG 31

Transitions / transversions = 1.64 (18 / 11)
Gap_init rate = 0.00 (0 / 270), avg. gap size = 0.00 (0 / 0)
```

图 16-8 RepeatMasker 鉴定的重复序列的两两比对结果的一个例子。顶部是一个简单的重复，底部是 Alu 的重复序列

```
Masked Sequence:
>gi|20548282:6973644-7073644 Homo sapiens chromosome 10
working draft sequence segment
AGTAAACAAAGGTTTTGXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXGCCTCTAAACTAA
AGGTTTTAATGAAGGAGAAGAAAAGCCTTGGGGGAAAAATGCTATTATTC
TGTTTGATTCACAAATTATGCATAATGGCACATGTGCTACCTTGCATGGA
TTATGAAGGCAAGCATTTTCACTTCAGTTTTGTAAGGTAGAGGTAAGGGG
CAGGAGAAGCTGATAATAGAGGATTAAGAAAAAAACTTGTAGAGTATATT
ATTATCAGCATAAACTTAGCAATCTGTTAATTAAATTTTGATCTGTTAAA
TGAGTTTAACAATATGTTGCATATATGCCACAGTAGTAATTTCTTCCCTT
GAAGGAGTGAACTTTACGGAAGTGATTCTGTTTATTGGCACTCAAAAGAG
AAGCACCTCAAATTAAAAAAAATTXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXCAATCAATTTAAATTTTTTCTTTTGTTTTACTGCTTGTTTTACATCAT
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXGTGGCCACTCTTCTATCACCCTAAAGCCAG
AAAATGTATGTGAAAGCACATTGCAGATGGCAAATACTGTCCCAGATTAT
TTTCATTTTTCAGCAATGACTGTAGTGTGGACGGAGCTGGAGAGAATGTG
AGAGAAGAAAGTGACATGACCTGCCCCCAGAACTGCCATTCATTTTACTT
GTCCATTATCACCATCTGTCATCCTCACCTGAAACTAGTATCAACGCAGG
TGCATTCAAATGCAAATAACAGAAAAGACTTACACCATCACGACATTTAT
TGTCTTTTTAAAAAGAAGTATGGAATCAGATGCTCGCAGTTTAGATCAGC
AATTTCACAAGGTTAGGGCTCTGGACTCTCATCTCTATCACTCTGCCCAA
GTTCACATCTCTACTGCTACAGTTCCAAGCCTCACATTCTAATAAGATTT
TTCTTCAAAACCAAGAAGACACCTTGCCACGGTCATGAATGAGCACTCCT
AATATGTTTCTTTTGTTTTATTGGAAAGTCCATTCTTTGGAAATGGAAAT
ATTTATAAGGAAACTCCACAGCACATGTCCCTTACATGTTATTGACTAGA
AATGGATATATATXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

图 16-9 FASTA 格式中 RepcatMasker 用 X 符号来隐藏重复性的序列，这些已经经过屏蔽处理的序列可以用来对数据库搜索以找寻它们的同源物

```
Summary:
==================================================
file name: RM2sequpload_8327
sequences:            1
total length:     99510 bp  (99510 bp excl N-runs)
GC level:         45.31 %
bases masked:     37952 bp ( 38.14 %)
==================================================
               number of      length   percentage
               elements*     occupied  of sequence
--------------------------------------------------
SINEs:               94      20982 bp     21.09 %
      ALUs           52      14798 bp     14.87 %
      MIRs           42       6184 bp      6.21 %

LINEs:               22       7643 bp      7.68 %
      LINE1           9       4876 bp      4.90 %
      LINE2          11       2508 bp      2.52 %
      L3/CR1          2        259 bp      0.26 %

LTR elements:        18       5980 bp      6.01 %
      MaLRs          11       3316 bp      3.33 %
      ERVL            4        975 bp      0.98 %
      ERV_classI      3       1689 bp      1.70 %
      ERV_classII     0          0 bp      0.00 %

DNA elements:        16       2268 bp      2.28 %
      MER1_type      11       1878 bp      1.89 %
      MER2_type       0          0 bp      0.00 %

Unclassified:         0          0 bp      0.00 %

Total interspersed repeats:  36873 bp     37.05 %


Small RNA:            0          0 bp      0.00 %

Satellites:           0          0 bp      0.00 %
Simple repeats:      16        758 bp      0.76 %
Low complexity:       7        330 bp      0.33 %
==================================================

* most repeats fragmented by insertions or deletions
  have been counted as one element


The sequence(s) were assumed to be of primate origin.
RepeatMasker version 07/07/2001 , default mode
run with cross_match version 0.990329
RepBase Update 6.3, vs 05152001
```

图 16-10 RepeatMasker 的输出包括了被确认的重复序列的种类。图中所示的结果是用包含了 *RBP4* 基因的人类 10 号染色体上的 10000bp 片段搜索得到。21%的序列包括了 *SINEs*，一共 37%的序列是散布的重复性序列

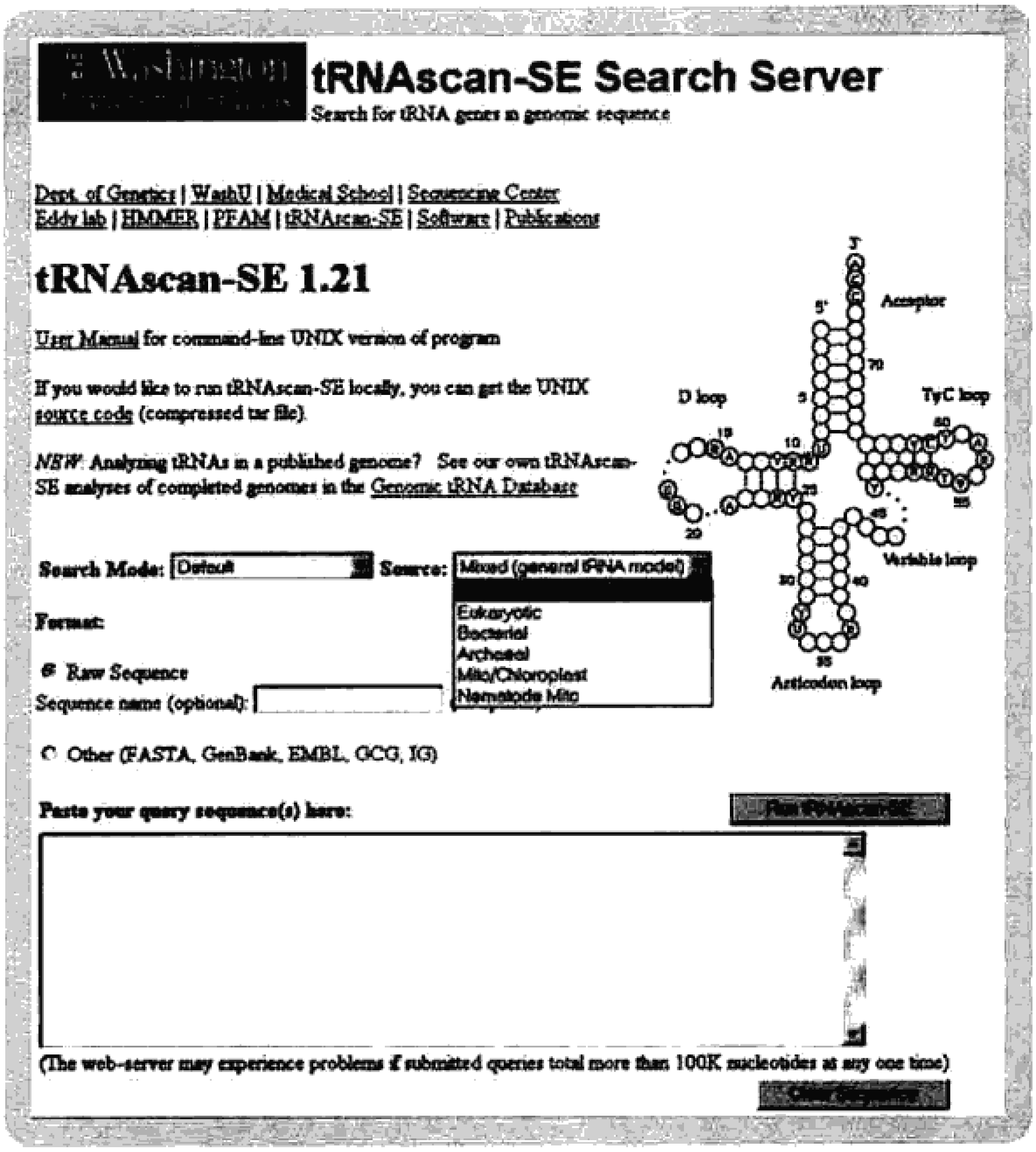


图 16-11 tRNAscan-SE 搜索网页服务器预测基因组 DNA 或者 RNA 序列中的 tRNA 分子。该程序也可以安装到本地机器以用于大型项目。tRNAscan-SE 并没有自己的算法，而是综合了 3 个不同的算法来找寻对应于 rRNA 特殊的序列的模式（如茎环结构）。tRNAscan-SE 也可以通过合适的二级结构的缺失特性来鉴定 tRNA 假基因

矩阵更加敏感，更加具有特异性。多个关于非编码 RNA 的数据库在表 16-8 中列出，例如 Rfam（Griffiths-Jones 等，2003）、RTNBase（Murthy 和 Rose，2003），以及非编码的调节 RNA 数据库（Szymansk，2003）。

表 16-8 非编码 RNA 数据库

数据库	描述	地址
Large-ribosomal-subunit database	大核糖体 RNA 结构	http://rrna.uia.ac.be/rrna/lsu/
Noncoding RNAs database	各种病毒 RNA 分类目录	http://biobases.ibch.poznan.pl/ncRNA/
Rfam	RNA 家族数据库	http://www.sanger.ac.uk/Software/Rfam/ 和 http://rfam.wustl.edu/
Ribosomal Database Project(RDP)	提供核糖体相关的数据分析，rRNA 引导的系统发生树，以及 rRNA 序列的比对与注释	http://rdp.cme.msu.edu/html/
RNABase	RNA 结构	http://www.rnabase.org
RNA Editing website	RNA 编辑的各种类型	http://www.rna.ucla.edu/
Small-ribosomal-subunit database	小核糖体 RNA 结构的数据库	http://rrna.uia.ac.be/rrna/ssu/
Small RNA database	小 RNA 被定义为不直接参与蛋白合成过程的 RNA	http://mbcr.bcm.tmc.edu/smallRNA/smallrna.html
tRNA sequences	550 条 tRNA 和 3704 个 tRNA 基因(through 1998)	http://www.uni-bayreuth.de/departments/biochemie/sprinzl/trna/index.html

16.2.6 在真核基因组中寻找基因

在真核基因组中寻找基因是比在原核生物中寻找基因复杂得多的一个问题（Burge 和 Karlin，1998；Mural，1999；Claverie，1997）。因为细菌的基因一般是与长的 ORF 联系在一起的，而真核生物的基因包括了内含子和外显子。外显子有多种（图 16-12）：

① 非编码的外显子即 DNA 序列的 5′非编码区和 3′非编码区。

② 包括启动位点的甲硫氨酸在内的起始外显子，直到第一个 5′剪切位点。

③ 内在的外显子，从 3′剪切位点开始，到 5′剪切位点结束。

④ 最后一个外显子，从 3′剪切位点到终止子。

⑤ 单外显子基因是没有内含子的，从起始密码子开始，到终止密码子结束（表 16-5）。

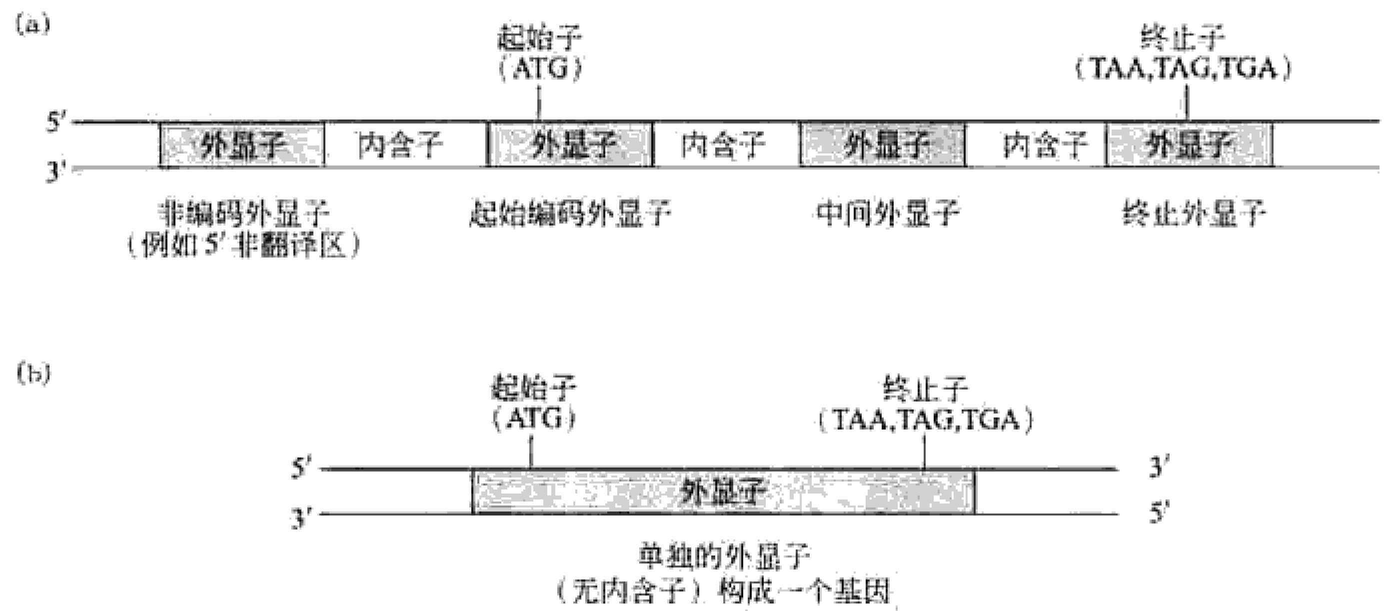


图 16-12 （a）真核基因预测算法区分不同种类的外显子，如那些处于非编码区域中的、包含起始密码子的起始的外显子，包含终止密码子的最后的外显子。这些外显子被整合进模型来预测基因。（b）某些情况下，基因只有一个外显子而没有内含子。外显子和内含子的边界通常是 GT/AG，但是这些结构仍然难以通过从头预测的方法来确定

人类基因的三分之一是可变剪切的。如果可变剪切形式的对应 ESTs 数据已经具备，这些序列可以被定位到外显子上。

除了具有内含子外，真核生物的基因占基因组的比例远远小于原核生物基因占基因组的比例。真核生物编码蛋白质的基因在线虫和昆虫基因组中只占 25%，而在人类和小鼠中只占不到 3%。

在真核生物中找寻编码蛋白质的基因的算法可以被分成两类：基于同源性的（也称为外在的）和基于算法的（也称为内在的）（Stein，2001）。这些算法在图 12-17 中被指出。基于同源的方法通常包括了对已表达的基因的比对（来自 cDNA 文库的 EST，见第 2 章和第 6 章）。在这种情况下，ESTs 可以帮助定义基因组 DNA 的外显子和内含子结构。因此，基于同源的方法通常是很成功的。另外一种基于同源鉴定基因的形式是把两个亲缘关系相近的物种基因组 DNA 进行比较（Morgenstern 等，2002；Novichkov 等，2001）。通过人类和河豚的基因组的比较，可能发现将近 1000 个新的人类基因（Hedges 和 Kumar，2002；Aparicio 等，2002）。

虽然用 EST 数据注释真核生物基因是非常有帮助的，但是这种方法也有一些限制条件。

① EST 数据的质量通常很低，只有一条链上的克隆，而且经常有序列上的错误。

② 高表达量的基因通常过于不成比例地出现，虽然一些 cDNA 文库已经经过归一化（第 6 章）。

③ EST 数据并不提供关于基因定位的信息。

内在性的方法也被广泛用于注释基因组 DNA。有相当一部分基因既没有已经鉴定的同源基因，也没有可以使用的 EST 数据，因此必须使用从头预测的方法来鉴定编码蛋白质的基因。我们在第 14 章讨论了用 GLIMMER 程序来预测原核生物的基因的方法。

许多基于网页的真核生物基因预测程序被列在表 16-9。其中包括了 Genescan（Burge 和 Karlin，1997）和 GRAIL。Rogic 等在 2001 年比较了 7 种程序的敏感度和特异性。这些算法利用编码的统计特性（例如编码的和非编码的区域的核酸序列的不同特性）和各种信号（例如，启动子元件、起始和终止密码子、剪切位点和多腺嘌呤位点），随着时间的推移而不断地改进。Makarov（2002）也比较了 7 种常用的真核基因预测程序的可用性和精确度。

表 16-9　真核基因 DNA 预测算法

程　序	描　述	地　址
AAT	基于网页的自动分析工具	http://genome.cs.mtu.edu/aat.html
FGeneH	用线性分割函数，预测外显子	http://genomic.sanger.ac.uk/gf/gf.html
FgeneSH	从头预测基因	http://www.softberry.com/berry.phtml
Gene Finder	人、鼠、拟南芥、酵母	http://argon.cshl.org/genefinder/
GeneParser2	鉴定基因组 DNA 中的编码区域	http://beagle.colorado.edu/%7Eecsnyder/GeneParser.html
Genie	基于 HMM 模型	http://www.cse.ucsc.edu/~dkulp/cgi-bin/genie
GenLang	模式识别系统，用计算机语言预测基因	http://www.cbil.upenn.edu/genlang/genlang-home.html
Genscan	基于 HMM，基于规则而不是同源性	http://genes.mit.edu/GENSCAN.html
GenTerpret	来自 RabbitHutch Biotechnology Corporation	http://www.rabbithutch.com/
GlimmerM	来自 TIGR	http://www.tigr.org/software/glimmerm/
GlimmerM web server	拟南芥，水稻，疟疾等	http://www.tigr.org/tdb/glimmerm/glmr_form.html
GRAIL	一个广泛使用的算法	http://compbio.ornl.gov/tools/index.html
MORGAN	一个基于决策树的系统，用于预测脊椎动物 DNA	http://www.tigr.org/~salzberg/morgan.html
PROCRUSTES	基因识别，基于分解的比对	http://www-hto.usc.edu/software/procrustes/index.html
VEIL	HMM 用于找寻脊椎动物 DNA 的基因	http://www.cs.jhu.edu/labs/compbio/veil.html
Xpound	一个概率模型用来检测编码区	http://bioweb.pasteur.fr/seqanal/interfaces/xpound-simple.html

注：1. HMM（hidden Markov model）表示隐马尔科夫模型。
2. 网址的总结位于 http://linkage.rockefeller.edu/wli/gene/。

虽然不同的软件使用的各种算法都不尽相同，但是它们都有一些共同的缺点，即过高的假阳性率以及不能准确预测出基因的完整结构。

Kapranov 等（2002）制造了与人类 21 号和 22 号染色体序列相关的 DNA 芯片，间距为 35 个核酸。该芯片包含了两个染色体的密度。他们从不同的人类细胞株中经杂交探针进行杂交，发现了 10 倍于已知基因的未知 RNA 表达产物。其中包含了许多未知的新基因。另外，他们也可能发现了大量的非编码的 RNA 和无功能的 RNA 种类。
Rogic 等（2001）的研究结果被放在 http://www.cs.ubc.ca/~rogic/evaluation/。
一个简单的预测编码蛋白质的基因的程序是 NCBI 的 ORF Finder。例如，输入 HIV-1 的基因组序列，该程序将预测出一系列 ORF。将预测结果同已经确认的 HIV-1 编码序列进行比较，发现许多假定的 ORF 实际上并没有被转录。

如果想知道真核生物的基因预测软件是如何工作的，请用橡树岭国家实验室（Oak Ridge National Laboratory）的基因组分析流程（图 16-13）。我们仍然使用人类 10 号染色体上的 10000bp 长度的 DNA 序列。在网站上选择合适的物种和算法后，输出的结果包括以下几个方面（图 16-14）。

① GrailEXP 程序的基因预测结果，结果是找到了 5 个基因，然后自动到蛋白质数据库和 Pfam 中用 blastp 进行搜索。

② Genscan 程序的基因预测结果，结果是找到了 3 个基因。

③ 对 CpG 岛的鉴定，通常与持家基因相关联（第 17 章）。

④ RepeatMasker 寻找重复序列，找到了 201 个（请与图 16-10 相比较）。

⑤ 编码转运 RNA 基因的预测（没有找到 tRNA 基因）。

⑥ Grail 程序寻找细菌人工染色体（BAC）配对，找到了 24 个结果。

⑦ E-PCR STSs（也就是 PCR 的标签序列），这些是基因组位置已经确定的由 NCBI 主页上 E-PCR 服务提供的短 DNA 序列（http://www.ncbi.nlm.nih.gov）。

因此，这些分析流程提供了一些有用的数据，包括了基于不同独立方法的基因预测。预测结果之间的不同表明了内在性预测方法所遇到的困难。GrailEXP 自带的工具 Perceval 把在 10000bp 中预测出的起始、中间和终止的外显子都进行了搜索（图 16-15）。对预测结果的评价既有非常高的也有处于最低可信度边缘的。对于 GrailEXP 预测出来的 5 个基因（图 16-16），

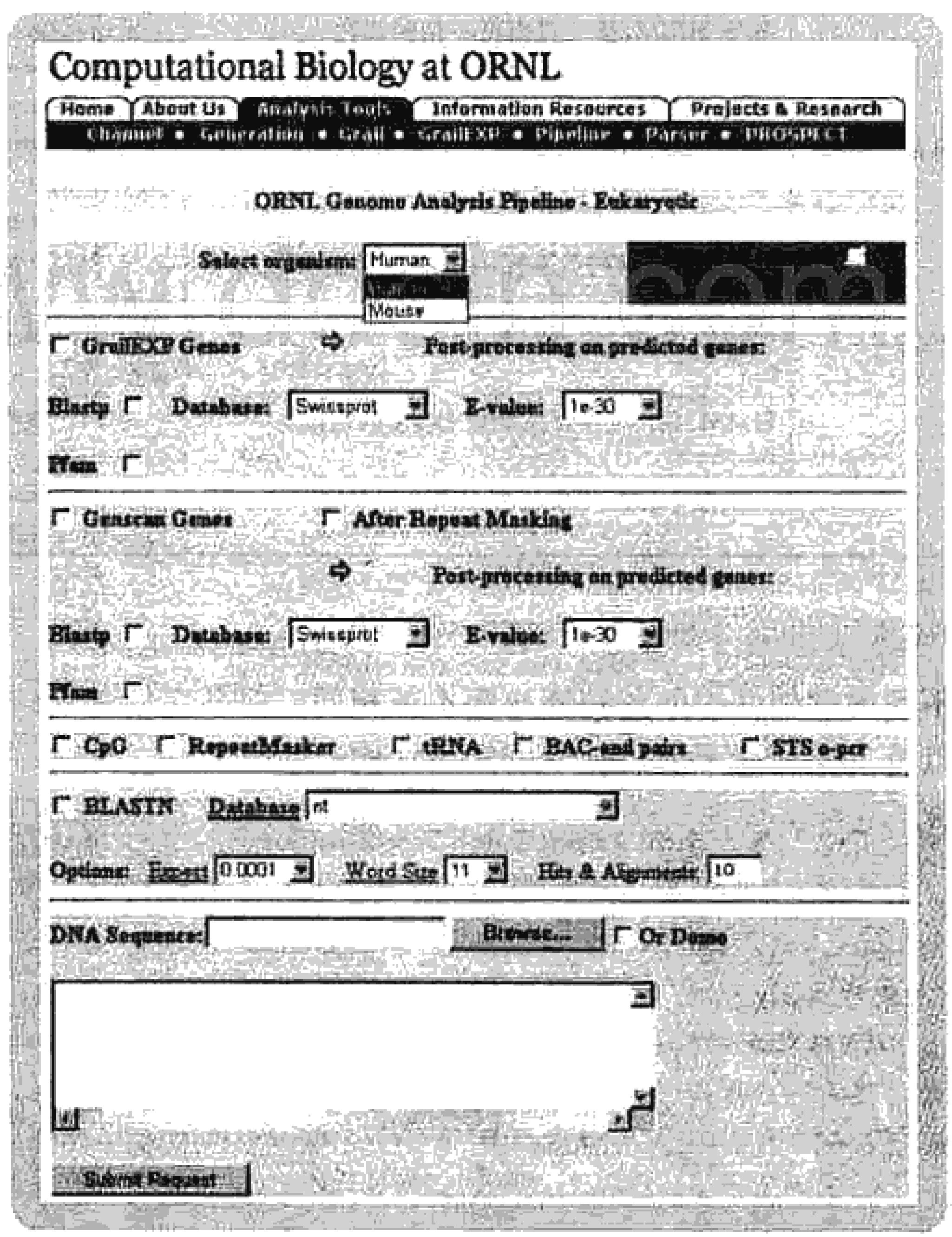


图 16-13 橡树岭国家实验室（ORNL）的基因组分析流程是一个基于网页的多物种基因组注释工具，包括了 GrailEXP 程序。将人类基因组 10 染色体上的包括 *RBP4* 在内的 10000bp 的序列作为该程序的输入。序列是从 NT_008769 获得（10921016bp）的。该流程的地址为 http://compbio.ornl.gov/tools/pipeline/

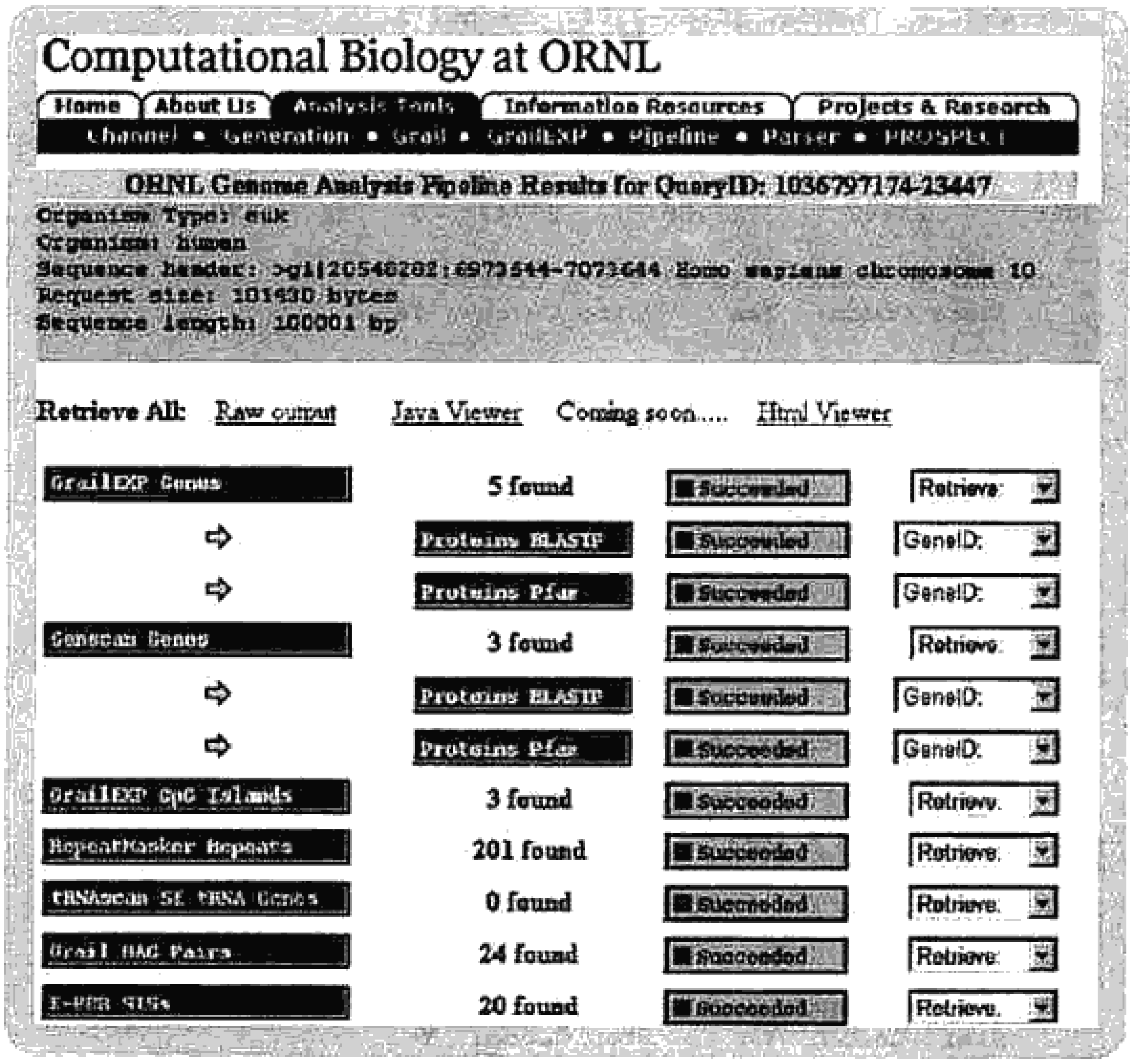


图 16-14 ORNL 基因组分析流程对一段人类基因组 DNA 片段的分析结果表明，GrailEXP 找到了 5 个基因而 Genscan 找到了 3 个基因，右边的下拉菜单提供了用 Blastp 和 Pfam 进行比对搜索的选项。这些算法正确地预测了 *RBP4* 基因，输出结果也包括了 RepeatMasker 找到的重复序列、tRNA 基因、BAC 克隆鉴定、来自 NCBI 的 e-PCR 库的序列标记位点（STSs）

PERCEVAL Exon Candidates (36 predicted)

Index	Std	Begin	End	Frm	Type	Len	Scr	Quality
10	+	35515	35557	0	Internal	43	61	Marginal
11	-	40500	40616	2	Internal	117	81	Good
12	+	43576	43965	0	Terminal	390	92	Excellent
13	-	50105	50391	1	Terminal	287	99	Excellent
14	-	52382	52436	2	Internal	55	53	Marginal
15	-	53575	53701	0	Internal	127	88	Good
16	-	54457	54518	0	Initial	62	44	Marginal
17	-	56741	56855	2	Terminal	115	90	Excellent
18	-	57023	57384	0	Initial	362	81	Good
19	-	57274	57384	0	Initial	111	85	Good
20	+	63682	63812	0	Internal	131	100	Excellent
21	+	69066	69462	2	Internal	397	88	Good
22	+	76988	77140	0	Internal	153	91	Excellent
23	+	77247	77336	0	Internal	90	90	Excellent
24	+	78288	78428	0	Internal	141	94	Excellent
25	+	81931	82005	0	Internal	75	93	Excellent
26	+	82996	83052	0	Internal	57	91	Excellent

图 16-15 GrailEXP 中的 PERCEVAL 程序预测了外显子的位置、类型（起始，中间和终止）、长度和得分，以及预测的可信度。对于来自人类染色体 10 的 10000bp 序列，该程序预测了 36 个外显子，图中给出了一些结果

我们可以通过简单地用 BLAST 进行搜索来判断其准确度。据此我们可以看到，预测出来的基因 3 和基因 4 实际上是 *RBP4* 的不同的外显子片段（图 16-17）。该程序也不能正确地建立起一个基因的模型。通过把预测的结果同实验鉴定的结果相比较，我们可以得出以上结论（图 16-18）。预测出来的基因 1、基因 2、基因 5 同染色体 10 上与 *RBP4* 邻近的几个基因相吻合。

```
Gene 1, Variant 1      Strand: +  Bounds: 18563-23742  Exons: 2
                            Start Codon:  Yes      Stop Codon:  Yes

 ---Index---- --------Exons-------- ---------CDS--------- -Ph- -Fr- -Len- -Scr-
 Promoter          14237     14350          ...       ...    .    .    114   51
 1.1.1             18563     18775     18563     18775    0    1    213   70
 1.1.2             23068     23742     23068     23742    0    0    675   88

Gene 2, Variant 1      Strand: +  Bounds: 43687-43965  Exons: 1
                            Start Codon:  Yes      Stop Codon:  Yes

 ---Index---- --------Exons-------- ---------CDS--------- -Ph- -Fr- -Len- -Scr-
 Promoter          39126     39349          ...       ...    .    .    224   70
 2.1.1             43687     43965     43687     43965    0    0    279   92

Gene 3, Variant 1      Strand: -  Bounds: 50105-50413  Exons: 1
                            Start Codon:  Yes      Stop Codon:  Yes

 ---Index---- --------Exons-------- ---------CDS--------- -Ph- -Fr- -Len- -Scr-
 PolyA             45824     45829          ...       ...    .    .      6   65
 3.1.1             50105     50413     50105     50413    0    1    309   97

Gene 4, Variant 1      Strand: -  Bounds: 56741-57384  Exons: 3
                            Start Codon:  Yes      Stop Codon:  Yes

 ---Index---- --------Exons-------- ---------CDS--------- -Ph- -Fr- -Len- -Scr-
 4.1.3             56741     56855     56741     56855    2    1    115   90
 4.1.2             57023     57159     57023     57159    0    2    137   71
 4.1.1             57274     57384     57274     57384    0    2    111   85

Gene 5, Variant 1      Strand: +  Bounds: 63682-97389  Exons: 14
                            Start Codon:  No      Stop Codon:  Yes

 ---Index---- --------Exons-------- ---------CDS--------- -Ph- -Fr- -Len- -Scr-
 5.1.1             63682     63812     63682     63812    0    0    131  100
 5.1.2             69066     69462     69066     69462    2    0    397   88
 5.1.3             76988     77140     76988     77140    0    1    153   91
 5.1.4             77247     77336     77247     77336    0    2     90   90
 5.1.5             78288     78428     78288     78428    0    2    141   94
 5.1.6             81931     82005     81931     82005    0    0     75   93
 5.1.7             82996     83052     82996     83052    0    0     57   91
 5.1.8             85614     85661     85614     85661    0    2     48   65
 5.1.9             91114     91267     91114     91267    0    0    154  100
 5.1.10            91920     91996     91920     91996    1    1     77   84
 5.1.11            93351     93419     93351     93419    0    2     69   97
 5.1.12            96426     96572     96426     96572    0    2    147   93
 5.1.13            96806     96913     96806     96913    0    1    108   93
 5.1.14            97276     97389     97276     97389    0    0    114  100
 PolyA             98736     98741          ...       ...    .    .      6   79
```

图 16-16 GrailEXP 程序预测了 5 个基因，其中基因 3 和基因 4 是已确认的 *RBP4* 基因的一部分

GrailEXP 程序套装的网络地址为 http://compbio.ornl.gov/grailexp/。请参见关于果蝇基因组的讨论，以了解关于注释基因组内编码蛋白质基因的难度的另外一个例子。

试图注释一个经典的真核基因组：对水稻的 *indica* 亚种和 *japonica* 亚种的基因组的努力充分说明了在基因组序列中发现编码蛋白质的基因的难度。Yu 等（2002）把他们得到的水稻基因组的草图提交到了 FGeneSH 服务器，得到了 75659 个预测出来的基因（表 16-9）。只有 53398 个预测结果是完全的（含有起始和终止外显子）：大约 7500 个结果只有起始外显子，11000 个结果只有终止外显子，3400 个结果既没有起始外显子也没有终止外显子。另外，他们发现外显子和内含子的边界通常都是不是很精确的。然而当水稻基因组序列的最终版本而不是草图被测定的时候，预测结果的数量有了明显的提高。Sasaki 等（2002）取得了水稻染色体 1 的最终版本的全序列（日本亚属），预测了 6756 个该

Gene 3 (predicted by GrailEXP) as query for a blastp nr search

```
>gi|5803139|ref|NP_006735.1|    retinol-binding protein 4, plasma precursor;
          retinol-binding protein 4, plasma; retinol-binding protein 4,
          interstitial [Homo sapiens]
 gi|72085|pir||VAHU   plasma retinol-binding protein precursor - human
 gi|35897|emb|CAA24959.1|   precursor RBP [Homo sapiens]
          Length = 199

 Score =  152 bits (385), Expect = 5e-38
 Identities = 71/72 (98%), Positives = 72/72 (99%)

Query: 8    DDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELC 67
          +DDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELC
Sbjct: 117 NDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELC 176

Query: 68   LARQYRLIVHNG 79
          LARQYRLIVHNG
Sbjct: 177 LARQYRLIVHNG 188
```

Gene 4 (predicted by GrailEXP) as query for a blastp nr search

```
 Score =  204 bits (518), Expect = 2e-53
 Identities = 105/120 (87%), Positives = 107/120 (88%), Gaps = 2/120 (1%)

Query: 1    MKWVWXXXXXXXXXGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV 60
          MKWV            AERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV
Sbjct: 1    MKWV--WALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV 58

Query: 61   AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGSE 120
          AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKG++
Sbjct: 59   AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND 118
```

图 16-17 GrailEXP 程序的输出结果包括了预测出来的蛋白质，并以 FASTA 格式提供给用户。当用它们对非冗余数据库进行搜索，基因 3 和基因 4 的蛋白质产物被匹配到 *RBP4* 上。但是预测结果中的第一个氨基酸与实验确认的 *RBP4* 并不匹配

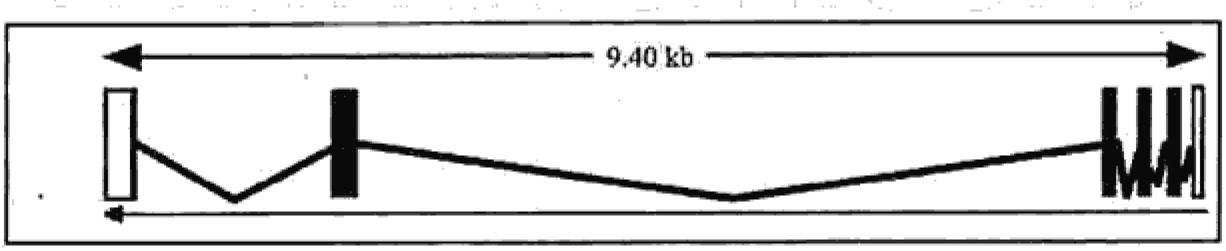


图 16-18 根据 Ensembl 的基因报告和来自 NCBI 的证据（通过 LocusLink 或者基因图浏览器），*RBP4* 基因的实际结构（来自 Ensembl 网页服务器，见第 17 章）包括了 6 个外显子。对于 GrailEXP 之类的算法来说，在没有实验数据（如全长 cDNA 序列）的条件下，想要准确地预测出全部的外显子并且把它们准确无误地组合成一个基因模型是很困难的

染色体上的基因。而在草图上的预测结果只有 4467 个基因。Sasaki 等（2002）认为，草图中存在的几千个序列缺口影响了对基因的精确预测。

16.2.7 真核生物中的蛋白质编码基因：新的悖论

对于 C 值矛盾的解释通常是不同真核生物基因组中非编码 DNA 的数量不同。然而一个新的矛盾又产生了：为什么虽然各种真核生物间的表型十分不同却有着相似大小

的蛋白质组？Claverie（2001）称之为 *N* 值矛盾（*N* 表示数字），而 Betran 和 Long（2002）称之为 *G* 值矛盾（*G* 代表基因的意思）。当我们在本章后半部分讨论不同真核生物的基因组时，我们将看到一些简单的生物（如蠕虫和线虫）含有大约 13000～20000 个编码蛋白质的基因，而植物、鱼类和人类的基因仅仅稍多一点（大约 25000～40000 个）（Harrison 等，2002）。为什么人类这种比蠕虫和昆虫复杂得多的生物种类只含有大约 2 倍于蠕虫和昆虫的基因的数目？高等的真核生物可能具有更复杂的基因调控模式，例如可变性剪切。

16.2.8 转录因子数据库和其他基因组 DNA 数据库

除了预测基因以外，预测基因组 DNA 序列中的其他特性也是很重要的（例如转录因子结合位点）。已经有许多用于这些目的的程序和储存基因组特性的数据库（表 16-10）。

表 16-10 基因组 DNA 启动子区特征鉴定软件

程 序	描 述	地 址
Ancient conserved untranslated DNA sequences(ACUTS)	分析真核生物基因（脊椎动物，昆虫和线虫）	http://pbil.univ-lyon1.fr/acuts/ACUTS.html
AliBaba2	在未知 DNA 序列中预测转录因子结合位点	http://www.gene-regulation.de/
Eukaryotic Promoter Database(EPD)	实验验证的 POL Ⅱ 启动子的注释信息，已被实验证实	http://www.epd.isb-sib.ch
FastM	转录调控单位模型（如启动子）	http://www.gene-regulation.de/
Gene Regulation	需要登录	http://www.gene-regulation.de
PlantProm	植物启动子数据库	http://mendel.cs.rhul.ac.uk
rSNP_Guide	转录因子结合位点数据库	http://util.bionet.nsc.ru/databases/rsnp.html
TRANSFAC	转录因子，结合位点，结合谱的数据库	http://transfac.gbf.de

一个包含了许多物种的内含子数据的数据库位于 http://www.introns.com/。
一个密码子使用数据库位于 http://kazusa.or.jp/codon/。

许多基因预测程序在它们的输出结果中给出 GC 含量（Knight 等，2001）。Gentles 和 Karlin（2001）比较了人类、酵母、拟南芥和果蝇基因组中的二核苷酸的组分和偏好性，发现了真核生物基因组内有各自不同的特定的模式。

16.2.9 真核基因组以染色体的形式组织起来

染色体是基因组 DNA 的组织形式。最近几十年来，利用比较基因组学的工具，已经可以估计不同物种的基因组内基因的数量、基因在染色体上的相对位置，以及染色体的一些其他特性，例如常染色质和异染色质的比例、GC 含量以及结构和调控相关的性质。利用已有的真核生物全基因组，我们现在可以深入探索染色体。

在每个物种中，倍数染色体的数目是一定的。我们可以使用许多数据库（如 NCBI，MIPS 和 SGD 等）提供的图形显示工具来探索酵母的 16 个染色体（第 15 章）。在人体内，染色体的数目是 46（也就是说，在所有的体细胞中，有 23 对染色体）。我们将在第 17 章和第 18 章中讨论用表意字符表示人类染色体的一些数据库。其他一些真核生物的表意字符可

以在表 16-11 的地址中找到。

表 16-11 染色体数据库

资源	注解	地址
Ensembl genome browser	人类、小鼠、大鼠、斑马鱼、河豚和蚊子图谱	http://www.ensembl.org/
The Unified Database for Human Genome Mapping	来自 the Weizmann Institute	http://bioinformatics.weizmann.ac.il/cgi-bin/udb/search_map_sbr.pl
Human chromosomespecific databases	来自 the U.K. HGMP Resource Centre	http://www.hgmp.mrc.ac.uk/GenomeWeb/human-gen-db-chromosomes.html
Ideogram Album	人、小鼠和马的图谱，来自 University of Washington	http://www.pathology.washington.edu/research/cytopages/
Human Chromosome Launchpad	来自 the Oak Ridge National Laboratory	http://www.ornl.gov/hgmis/launchpad

当我们探索一系列被测序的真核生物基因组时，描述染色体的结构和内容是十分有用的。用 Wright 染色法染色后的人类中期染色体组型见图 16-19。一系列的染色实验产生了许多染色体上的带型。它们包括 Q 带型（用奎纳克林芥末黄及其衍生物染色）和 G 带型（用 Giemsa 染料，Wright 染色法就是基于此种染料）。这些染料把染色体的全长都染上颜色，形成一个特有的带型模式。一个带型就是染色体的一部分，它因为同周围的片段相比呈现明亮或者黯淡而区分开来。



图 16-19 人类染色体组型的一个例子。染色体用 Wright 染料染色。着丝粒位于染色体的凹痕位置处（如箭头 A 处）。图中所示的是一个染色体 11q 端粒附近（箭头 B）半合子缺失的人的带型

真核生物染色体有几种主要的特性。最主要的特征就是端粒（染色体的末端）和着丝粒（一个不能被许多染料着色的区域）。

我们也许会认为染色体是各个物种中用于定义基因组的不变的物质。但是，在时间的长河（几百万年）中，它们也是在代与代之间，甚至是个体的生活周期内不断变化的。大量的细胞内源性的变化不断发生，而且许多变化会导致不正常的结果。

① 整个基因组的复制事件（同源多倍性）有时会发生。这种事件发生在酵母（第 15 章）、植物（如下文）和其他许多种系中。

一个表意符号是用来表示一个染色体组型的符号。染色体组型是细胞中期当每条染色体都是一对姐妹染色单体时的一个染色体图像（通常是照片）。我们用一组有顺序的数字来表示染色体，染色体的短臂朝上。较短的一条臂被称为p臂是因为petit在法语中是小的意思，而q臂（长臂）因为在p字母后面而被命名。

人们通常研究处于细胞中期的染色体，因为这时染色体最厚最稠密。通常样本是在血液或者羊水中收集。染色体通过染料染色或者荧光原位杂交（FISH）而变得可见。

② 不同的两个物种的基因组有可能合并在一起形成一个新的物种（异源多倍性）（Hall 等，2002）。这种现象在许多植物（Comai，2000）、动物和真菌中被发现。例如，植物 *Arabidopisis suecica* 是拟南芥（*A. thaliana*）和 *Cardaminopsis aerenosa* 基因组衍生而来（Lee 和 Chen，2001；Lewis 和 Pikaard，2001）。异源多倍性的另外一个例子是骡子，它是公驴（*Equus asinus*，$2n=62$）和母马（*Equus caballus*，$2n=64$）的杂交后代。骡子不能繁殖是因为它们是不育的（不能产生有功能的单倍体配子）（见 Ohno，1970）。

③ 等臂染色体的着丝粒处于染色体的中央。近端着丝的染色体的着丝粒离端粒很近。在小鼠中，所有的染色体都是近端着丝的。在人体内，近端着丝的染色体有13号、14号、15号、21号和22号染色体。这些近端着丝粒通常发生 Robertsonian 移位而导致两个着丝粒融合。

④ 一个个体可能会含有某个染色体的多余拷贝。例如 Down's 症就是因为有3条21号染色体而导致的。我们将在第18章中讨论这种紊乱。

⑤ 单性生殖的二倍性也可能出现，所有的同源染色体都是来自于同一个父亲。这些也跟人类疾病相关（Kotzot，2001）。

⑥ 染色体可以融合。人类2号染色体（人体内第二大染色体），是由远古巨猿的两个染色体衍生而来（黑猩猩染色体12和染色体13）（Ijdo 等，1991a；Martin 等，2002；Fan 等，2002）。位于着丝粒附近的人类的2q13带型，包括了按头对头方式排列的端粒重复序列。多于50个的位于染色体内部的端粒已经被报道（Azzalin 等，2001）。

你可以在 NCBI 的人类孟德尔遗传在线资源（OMIM）站点（条目编号306700）找到关于血友病的更多资料。我们将在第18章描述 OMIM。

⑦ 染色体的区域可能被倒置。例如，蚊子 *Anopheles gambiae* 有5种不同的亚型，它们的2号染色体上有不同种类的同臂内倒置序列（Holt 等，2002）。在人体内，一个基因的倒置可能导致严重的血友病（Antonarakis 等，1995）。

⑧ 染色体的一部分可能被删除。删除可能发生在末端也可能发生在中部。图16-19（箭头B）显示了染色体11q的末端删除的一个例子。

⑨ 片段性的复制经常发生，上面已经进行了讨论。

⑩ 来自任何真核生物物种的不同个体的染色体可能在长度、数目和异染色质片段的位置等各个方面都有所不同。

11q被删除会导致三角头畸形（头部形状类似三角形），如鲤鱼一样的尖嘴巴和心脏功能缺陷（Jones，1997）。

⑪ 脆性的位点经常会断裂从而导致染色体断裂。这些脆性位点可以通过孟德尔显性方式遗传给下一代。

⑫ 至少有一些真核生物的染色体有染色质减少的现象，它是一种发育过程中 DNA 计划性重排的形式。值得注意的是，体细胞内的染色体可能会碎裂，然后失去一些染色体物质。因此，相对于种系细胞来说体细胞可能含有不同的结构组织和更少的基因数目。

染色质缩减现象可以表现出一种不同寻常的基因沉默机制（Muller 和 Tobler，2000）。这种现象在至少 10 种蠕虫中被发现，包括马体内的寄生虫 *Parascaris univalens*（也称为 *Ascaris megalocephala*）和猪体内的寄生虫 *Ascaris suum*。

16.2.10 真核生物 DNA 比较：PipMaker 和 VISTA

比较基因组学是一种注释和理解来自多种物种的基因组 DNA 序列的有力方法。当我们比较分化时间较近的物种（如人类和黑猩猩于 500 万年前分化）或者分化时间较远的物种时（如蚊子和果蝇于 2.5 亿年前分化，Zdobnov 等，2002），把多个物种的序列比对起来以确定保守区域的方法是有用的。这样的分析可以提供关于编码蛋白质的基因的存在和进化以及其他 DNA 特性的非常有价值的信息，同样也能提供关于染色体进化的信息。

拥有一个共同祖先并且具有共同的功能，但是来自不同的物种的基因称为直系同源基因（第 3 章）。我们希望通过比较来自 2 个或者更多物种的基因组序列，分析每个物种中含有直系同源基因的那些区域。这些区域被称为保守的同线性区域。同线性表明在同一个染色体上出现了 2 个或者多个基因座，不管它们是不是连锁状态。这个定义描述了单个物种中一个染色体上的基因排列情况。而“保守同线性”描述了直系同源基因之间（例如两个物种间）的共线性现象。例如，人类染色体 10 上基因 *RBP4* 和 *CYP26A1* 和小鼠染色体 19 上的同源基因就体现出了保守共线性现象（图 16-27）。

为了鉴定保守共线性的区域，或者更大的并不包含有编码蛋白质的基因的 DNA 序列区域，需要对基因组 DNA 进行两两比对和多序列比对。我们已经讨论过原核生物中的这类方法（第 14 章），我们也讨论了类似 BLAT 的能将大段的 DNA 同数据库里的基因组 DNA 进行比较的算法（第 5 章）。

> 同线性来自于希腊语，原意为“同一条线”或者“同一条带子”。一个常见的错误是把共线性而不是保守共线性的定义作为直系同源基因的性质（Passarge 等，1999）。

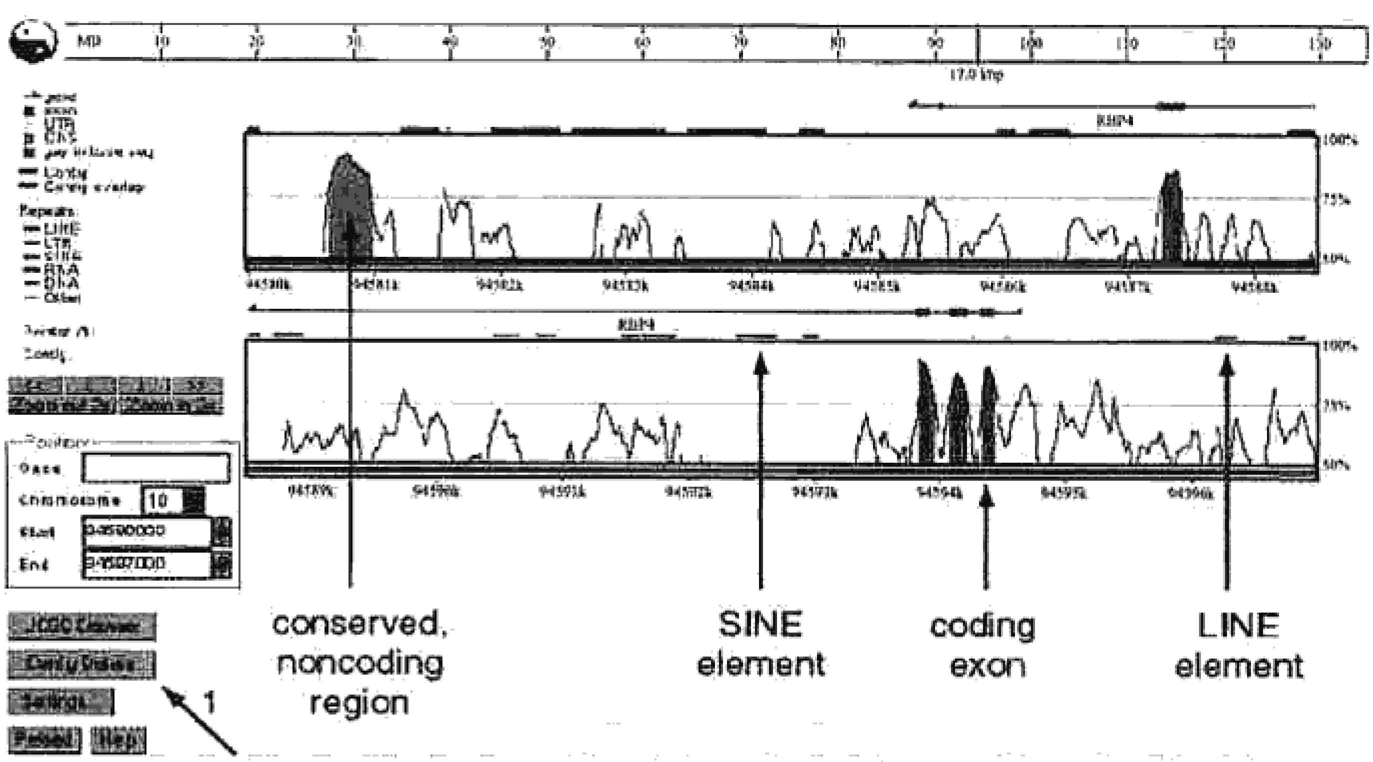
两个比较真核生物基因组 DNA 的经典程序是 PipMaker（Schwartz 等，2000）和 VISTA（Mayor 等，2000；Frazer，2003）。每个程序的目的都是把大段的序列（例如从几千到几百万碱基对）进行比对并且把保守的区域（外显子和调控序列区域）以及大段的基因组序列变化（例如倒置，重排，复制）显示出来。掌握保守序列的排列顺序和方向特性是很重要的。

作为一个执行基因组比对的例子，请到 Berkeley 的 VISTA 浏览器然后敲入 *RBP4*。运行结果显示了人类染色体 10 上的序列（*x* 轴）和人与小鼠之间的序列相似性（*y* 轴）（图 16-20）。灰色底纹部分显示诸如可能的编码区域、保守的非编码区域和重复 DNA 等元件。PipMaker 和 MultiPipMaker 提供了类似的输出结果。图 16-21 给出了另外一个例子：将人类染色体 10 上的 100000 核酸序列和小鼠 19 号染色体上的相应序列以 FASTA 格式作为输入，输出结果包括了一个点图（此处略）、一个匹配区域的概况 [图 16-21(a)]、一个两两比对的结果 [图 16-21(b)]，以及结果统计 [图 16-21(c)]。类似 VISTA 和 PipMaker 这样的程序将在基因组分析中发挥越来越大的作用。

> 内部健康测序中心国家研究所（NISC）目前正在进行多个物种的保守共线性的基因组区域的 BAC contig 序列测序工作。被研究的物种包括了人类、狒狒、黑猩猩、牛、小鼠、大鼠、狗、猫、鸡、斑马鱼和河豚（详见 http://www.nisc.nih.gov）。

PipMaker 和 MultiPipMaker 可以在如下网址下载到：http://bio.cse.psu.edu/pipmaker/（Pip 指 "percent identify plot"）。VISTA（比对可视化工具）在 http://www-gsd.lbl.gov/vista/。mVISTA（main VISTA）是一个用来可视化比对结果的工具，而 rVISTA（regulatory VISTA）被用来比对转录因子结合位点。AVID 是一个比对算法，它被 VISTA 所采用（Bray 等，2003）。Berkeley 基因组流程分析包括了一个 VISTA 浏览器（http://pipeline.lbl.gov/），它允许用户进行人类-小鼠、小鼠-大鼠、人类-大鼠的基因组 DNA 序列比较。

(a)



(b)

Criteria: 70% identity over 100 bp

*************** Conserved Regions ***************

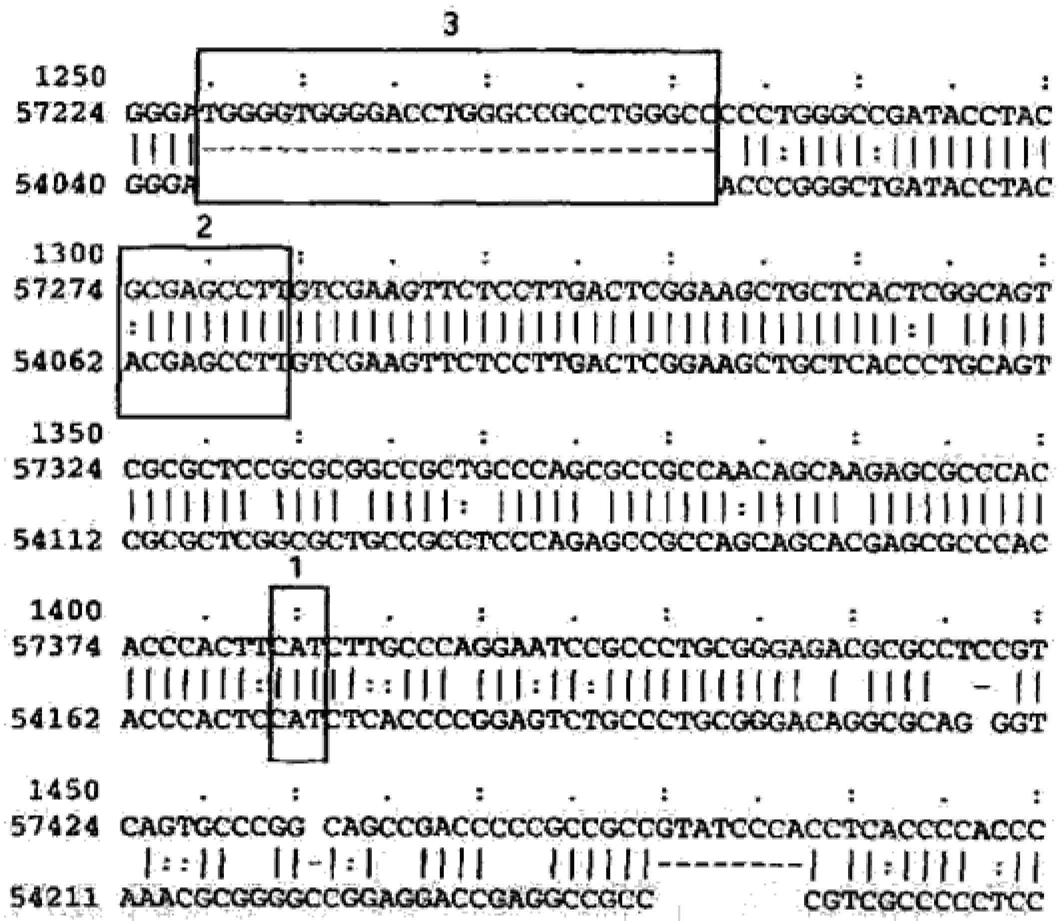
94585364	to	94585486	=	129bp	at	69.80%	UTR
94594441	to	94594458	=	18bp	at	72.20%	UTR
94594583	to	94594652	=	70bp	at	81.40%	UTR
94587237	to	94587445	=	209bp	at	83.30%	exon
94593805	to	94593910	=	106bp	at	91.50%	exon
94594080	to	94594215	=	136bp	at	86.80%	exon
94594331	to	94594440	=	110bp	at	90.90%	exon
94589637	to	94589864	=	229bp	at	72.50%	noncoding
94589940	to	94590050	=	112bp	at	69.60%	noncoding
94590435	to	94590544	=	111bp	at	73.00%	noncoding
94591250	to	94591381	=	133bp	at	73.70%	noncoding
94593365	to	94593457	=	93bp	at	72.00%	noncoding

图 16-20 用于比对基因组 DNA 序列的 VISTA 提供了一个基于网页的浏览器，用户可以提交文本或者 DNA 序列（多达 300000 个碱基）。(a) 用 *RBP4* 的文本搜索结果。x 轴显示的是人类染色体 10 上的核酸位置，y 轴显示的是人类和小鼠间序列的相似性。许多的编码和非编码特性被指出。这里也有一个超链接连接到 UCSC 基因组浏览器（在第 17 章中描述）。点击 Contig Details（箭头 1），关于比对区域的概况将被显示出来 (b)

(a)

```
Aligned to ">gi|20889106:4335093-4435093 Mus musculus WGS supercontig
Mm19_WIFeb01_321":
:
1040-1099 <--> 36101-36160 73% (60 nt)
1100-1128 <--> 36168-36196 55% (29 nt)
:
1228-1239 <--> 9578-9589 92% (12 nt)
1240-1255 <--> 9598-9613 44% (16 nt)
1257-1296 <--> 9614-9653 68% (40 nt)
1299-1301 <--> 9654-9656 100% (3 nt)
1311-1328 <--> 9657-9674 61% (18 nt)
1329-1505 <--> 9676-9852 62% (177 nt)
1508-1524 <--> 9853-9869 76% (17 nt)
:
```

(b)



(c)

```
Local Alignment Number 317
Similarity Score:  92303
Match Percentage:  62 %
Number of Matches:  1985
Number of Mismatches:  752
Total Length of Gaps:  456
Begins at (55993,52948) and Ends at (59050,55819)
```

图 16-21 PipMaker 程序对大段的基因组 DNA 序列进行两两比对：分别来自人类染色体 10 和染色体 19（gi 20889106）的一部分。比对包括了表格形式的概况（a），比对［一部分在（b）中给出］和一个综述（c）。这个区域包括了底部链的 *RBP4* 基因。起始密码子（ATG）出现在底部链的 5′TAC3′（方框 1）。可以发现人类 *RBP4* 外显子 2 的最后 9 个核酸（5′AAGGCTCGC3′编码了亮氨酸-丙氨酸-精氨酸）是作为倒置的补足物（5′GCGAGCCTT3′）（方框 2）。在它们的下游区域——小鼠基因组的一个内含子中出现了 gap（方框 3）。本图中人类的序列可以同图 6-4 中的相比较

16.3 真核生物基因组个例

16.3.1 引言

我们转向探索真核生物基因组。对所有基因组的分析主要包括以下几个问题：①描述每

个染色体的全部基因组序列；②对 DNA 进行注释以确定非编码 DNA；③鉴定编码蛋白质的基因和其他非编码的基因。进一步的问题包括：①染色体结构的分析，例如复制区域和删除区域的分析；②对相关物种的基因组进行比较基因组学分析以试图重新构建物种的分子进化树；③将基因组 DNA 序列同物种的表型相联系的规则。这些表型包括了对环境的适应、进化、代谢过程、生长、发育、维持内稳态和繁殖。

protozoan（原核生物）这个词来源于希腊文的单词 proto（早期的意思）和 zoion（动物的意思）。与此形成对比的是，metazoan（动物）这个词来自于希腊文的 meta（后期的意思，处于一个发展阶段的后期）和 zoion。

对于许多真核生物，从原生生物（例如 plasmodium）到致病性真菌再到寄生的蠕虫，我们也还想了解这些生物致病的基因基础以及我们如何对付它们。到目前为止，没有疫苗能用来对付寄生在人身上致病的真核生物，包括原核生物（例如锥体虫）和肠虫（寄生性的蠕虫）。对整个基因组序列的了解将为了解哪些抗原是疫苗开发和药物设计的可行靶标提供线索。例如，预测出来的表面蛋白可以在细菌中表达，然后用来激发小鼠的免疫系统以开发潜在的疫苗（Fraser 等，2000）。

当我们开始讨论真核生物基因组时，我们重新回顾图 16-1 中的种系发生树。我们将从这棵树的底部到顶部描述基因组测序项目。

一个真核生物种系发生树对我们理解物种的进化过程和当今生命的多样性都是必要的。进化树的重建是建立在分子序列的基础上，通常使用核糖体 RNA 的小亚基序列，因为它广泛存在于所有的生命形式中，而且许多位点都含有丰富的种系发生信息（Van de Peer 等，2000）。我们在图 12-2 中给出了这样的树的一个例子，但是目前没有一个公认的惟一的构建进化树的最佳方法（框 16-1 和第 11 章）。

框 16-1 不同的种系树

需要指出的是，构建的许多种系树是互不一致的。导致相互矛盾的原因主要有以下 3 个方面（Philippe 和 Laurent，1998）：

① 基因的复制和随之而来的随机基因丢失会在种系树的构建过程中会受到人为因素影响。这类事件在酵母以及其他真核生物（如植物和鱼类）的整个基因组水平上发生过（第 15 章）。

② 横向的基因转移给对种系树的解释带来干扰（第 14 章）。

③ 对长分支吸引的技术因素影响会对种系分析产生干扰。最长的分支通常会被聚集在同一个组，而不管真正的树的结构如何（图 11-23）。对同一个分子上不同位点的替换速率的差异的解释也是必要的。Reyes 等（2000）在他们的啮齿类动物的种系树构建中考虑了这个问题。

研究者们通常使用同时考虑多个蛋白（或者核酸）的序列的策略来解决许多问题。例如，图 16-1 中的树是基于 4 个蛋白。Wang 等（1999）用 75 个基因构建了他们的动植物和真菌的种系树（下文将要讨论）；Kumar 和 Hedges（1998）研究了 207 个脊椎动物中的 658 个基因。另一种策略是基于基因组内容和融合事件的，多个研究小组利用这种策略构建进化树（Snel 等，1999；Stechmann 和 Cavalier-Smith，2002）。

16.3.2 位于树底层的原生生物

真核生物包括了分支众多的原生生物，从 *diplomonadida*（如 *Giardia* 和 *Trihomonas*），*discicristat*（如 *Euglena*，*Leishmania* 和 *Trypanosoma*），*alveolata*（如 *Toxoplasma* 和 *Plasmodium*）以及 *heterokonta*（图 16-1）。

存在有力的证据证明现在的大多数真核生物都含有的线粒体基因是从 α-proteobacterium（见第 14 章）衍生而来。以前人们认为处于种系树分支深处的生物（如 *Giardia* 和 *Trichomonas*）缺少线粒体。曾经认为它们是由 α-proteobacterium 人侵之前的真核生物进化而来。但是，最近对 *Giardia*、*Trichomonas* 和小孢子虫（如 *Trachipleistophora bominis*）的分析表明了线粒体基因的存在（Embley 和 Hit，1998；Williams 等，2002；Lloyd 和 Harris，2002）。

Giardia lamblia（也称 *Giardia intestinalis*，贾第鞭毛虫）是一种原生生物，出生于水中，寄生在人和鸟的肠道里（Adam，2001）。它是贾第鞭毛虫病的引发者，这种病是北美地区最常见的非细菌性腹泻的病原。类似于其他的一些单细胞原生生物，贾第鞭毛虫不但缺少线粒体，也缺少过氧物酶体（负责脂肪酸的氧化）和核仁。因此贾第鞭毛虫的基因组反映了真核细胞出现的早期的对环境的适应。

贾第鞭毛虫的基因组长度大约是 12Mb（Adam，2001），对它的全基因组现在正在用鸟枪法测序之中（McArthur 等，2000）。每个细胞含有两个形态上的核，每个核中含有 5 条染色体，大小从 0.7～3Mb。一个在线的数据库描述了被注释的贾第鞭毛虫基因组的 ORF。

> diplomonadida 也被称为 diplomonads。这个目包括了六鞭科，而六鞭科包括了贾第鞭毛虫类（描述如下）。
> 小孢子虫（如 *Encephalitozoon*）过去被分类成分支很深的真核生物。后来对 *E. cuniculi* 的基因组的深入分析发现，这种小孢子寄生生物跟真菌的亲缘关系很近（第 15 章和图 16-1）。
> 贾第鞭毛虫是第一个用显微镜发现的人类寄生原生生物，发现者是 Antony van Leeuwenhoek（1681）。
> 美国食品和药物监督管理局提供了关于贾第鞭毛虫的信息，网址是 http://vm.cfsan.fda.gov/~mow/chap22.html。
> 缺少过氧化物酶体的生物可以为我们研究脂肪酸代谢途径和其他代谢途径提供信息，这些可以帮助我们了解影响这些细胞器的人类疾病。最常见的影响过氧化物酶体的遗传缺陷是肾上腺脑白质营养不良，由 *ABCD1* 基因的突变引起（RefSeq 编码 NM_000033）。贾第鞭毛虫含有这个基因吗？

当我们考虑各种真核生物的基因组时，一个普遍的现象是它们都含有非常丰富的转座子元件，几乎占到了整个基因组的一半（第 17 章）并引发了大量的基因组重排事件。因此为了发现它们的起源和功能，人们对那些没有转座元件的真核生物非常感兴趣。贾第鞭毛虫提供了这样一个例子。Arkhipova 和 Meselson（2000）研究了 24 种真核生物的两种主要的可移动元件（反转录转座子反转录酶和 DNA 转座子）。他们发现这两种元件存在于所有的物种中，只有一个物种除外，那就是蛭形轮虫，一种无性生殖的动物（见表 16-22）。有害的可移动元件可以在有性生殖的物种中保留下来，但是不能在无性生殖的物种中遗传，因为存在着对活性元件的强大的选择压。Arkhipova 和 Morrison（2001）对无性生殖的贾第鞭毛虫的进一步研究发现了 3 种反转录转座子家族。其中一个已经失去活性而另外两个是端粒。这个转座子所在的位置可以在编码蛋白质的基因和端粒之间提供一个缓冲区。这些元件可能对贾第鞭毛虫染色体长度可根据环境条件而变化的特点有贡献。例如，染色体 1 可以在 1.1～1.9Mb 范围内变化（Pardue 等，2001）。

> 贾第鞭毛虫基因组计划的网址是 http://www.mbl.edu/Giardia（McArthur 等，2000）。
> 你可以在 GenBank 中找出贾第鞭毛虫铁氧化还原蛋白基因（DNA 编号 AF393829）。为了找出内含子，用 BLAST（第 3 章）比较蛋白质和基因组 DNA 序列。

另外一个基本的问题是真核生物基因组中内含子的起源。内含子在真核生物的“顶部”

(动物、植物和真菌)中比较常见。但是内含子在较早期的原生生物中是否存在还处于争议之中(Johnson, 2002)。尚未在 parabasalids(如 *Trihomonas*)中发现内含子。Nixon 等(2002)鉴定出了位于被认为编码了铁氧化还原蛋白的一个基因中的 35bp 长的内含子。Simpson 等(2002)接着在 *Carpediemonas membranifera*(一种被认为跟贾第鞭毛虫亲缘关系很近的真核生物)中鉴定出多个内含子。这些研究表明,如果内含子是由于真核生物对环境的适应而产生的,那么它应该在进化的早期就已经存在,而且很可能存在于真核生物的共同祖先中。

16.3.3 单细胞病原体的基因组:锥体虫和利什曼虫

原生生物锥体虫类大约有 20 个物种(Donelson, 1996)。其中的 2 个物种对人类来说是致病的(Cox, 2002)。Trypanosoma brucei 亚种引起多种昏睡症,这种病症感染了非洲的数十万人口。*Trypanosoma cruzi* 引起 Chagas 症,在南部和中部非洲非常流行。锥体虫也能感染牲畜,使得它们的危害更大。舌蝇和其他昆虫是锥体虫的传播者。

非洲锥体虫基因组网络正在为测定锥体虫基因组序列而努力(表 16-12)。*T. brucei* 的基因组大小为 35Mb,虽然不同的样本中差异可达到 25%(El-Sayed 等, 2000)。*T. brucei* 至少有 11 对大的二倍性的核染色体(大小从 1~6Mb 不等);另外,还有数目不定的大小中等的染色体(200~900kb),还有大约 100 条线性的小染色体 DNA 分子(50~150kb)。有些小染色体包括了一种 177bp 长的重复序列,可以占到整个染色体的 90%以上(El-Sayed 等, 2000)。

表 16-12 锥体虫基因组网络资源

资源	注解	地址
The *Trypanosoma brucei* Genome Network *Trypanosoma brucei* omni Blast Server	由 Wellcome Trust Sanger Institute 资助	http://parsun1.path.cam.ac.uk/ http://www.sanger.ac.uk/Projects/T_brucei/Toolkit/blast_server.shtml
Trypanosoma cruzi Genome Initiative Information Server	来自 the Oswaldu Cruz Institute, Brazil	http://www.dbbm.fiocruz.br/TcruziDB/index.html

舌蝇是一种吸食脊椎动物血液的昆虫,为了获取血液之外的营养物质,舌蝇携带了两种细菌:*Wigglesworthia glossinidia* 和 *Sodalis glossinidius*。*W. glossinidia* 的基因组最近已经被测序(见 RefSeq 编号 NC_004344)。与另一种细胞内细菌相似(第 14 章),它只有一个大小 700000bp 的基因组(Akman 等, 2002)。

锥体虫的另一个明显特性是它含有许多环状的线粒体 DNA(称为动基体 DNA)。几千个动基体 DNA 联动形成了一个保护屏障(Shapiro 和 Englund, 1995)。动基体 DNA 形成几个大的环(几十个拷贝)和小的环(几千个拷贝)。其中包括了广泛存在的作为复制起点的 12 个核酸的小环状序列(Morris 等, 2001)。

利什曼虫是另一种致死的原生生物寄生虫(图 16-1)。13 种不同种类的利什曼虫可以导致利什曼病,到目前为止还没有有效的疫苗和药物可治疗此病。这些不同种类的利什曼虫含有 34~36 个染色体(Myler 等, 2000)。

利什曼虫的全基因组含有 36 个染色体,大约 34Mb(从 0.3~2.5Mb)。利什曼虫基因组网络负责对其测序的工作(表 16-13)。染色体 1 的序列已经被测出(最小的染色体),发现其组织形式非常特别(Myler 等, 1999)。前 29 个基因(从左端粒开始)都是从同一条单链上转录的,其他的 50 个基因都是从另外一条链上转录

的。这种情况以前从来没有在真核和原核生物中发现过。染色体 1 含有一个 257kb 长的区域，其中排列 79 个编码蛋白质的基因（每个基因约 3200bp）。Myler 等（1999）估计全基因组内大概有 9800 个基因。

表 16-13 利什曼虫基因组网络资源

资 源	注 解	地 址
The *Leishmania major* Friedlin Genome Project	位于 the Wellcome Trust Sanger Institute	http://www.sanger.ac.uk/Projects/L_major/
Seattle Biomedical Research Institute (SBRI)	多种传染疾病相关信息	http://www.genome.sbri.org/
The European *Leishmania major* Friedlin Genome Sequencing Consortium	多个实验室	http://www.sanger.ac.uk/Projects/L_major/EUseqlabs.shtml

16.3.4 疟疾致病体——疟原虫

疟疾每年使 270 万人口死亡，大多数都是非洲的儿童。全球每年新增大约 5 亿新感染的人口。这是由于疟原虫感染导致。在总共的 120 种变形体种类中只有 4 种能感染人类：*P. falciparum*，*P. vivax*、*P. ovale* 和 *P. malariae*。非洲疟疾的主要载体是蚊子。

疟原虫的生活方式很复杂，这给制造有效的疫苗带来了困难（Cowman 和 Crabb，2002；Wirth，2002；Long 和 Hoffman，2002）。疟原虫居住在蚊子的唾液腺和内脏内。当蚊子咬人的时候，就把疟原虫的孢子体注入人体感染肝脏。疟原虫成熟后变为裂殖体，然后通过宿主的细胞受体侵入人体内的红细胞，在红细胞内形成营养体。一些裂殖体转变成为配子，当蚊子吸食人血的时候附着到蚊子体内。测定疟原虫的基因组序列的一个目的就是找出基因组中那些在生活周期中的不同阶段起作用的基因产物，从而为药物的研究和疫苗开发提供靶标。

可以从世界卫生组织的网站 http://www.who.int/tdr/diseases/tryp/default.htm 获得关于睡虫病的更多信息。

见问题 3，有一个关于锥体虫的小环结合蛋白的练习。参见 GenBank 编号 M94286，以获得关于大环序列及其编码的基因的一个例子。

世界卫生组织提供了关于利什曼病的信息，见 http://www.who.int/inf-fs/en/fact116.html 和 http://www.who.int/tdr/diseases/leish/default.htm。

Alveolates 包括了 3 个分支的原生生物：①Apicomplexa 门，包括了例如疟原虫这样的寄生虫；②纤毛类，包括了草履虫和四膜虫；③腰鞭毛虫，包括了一种能危害甲壳类动物的 Alxandrium。

Apicomplexa 这个名字来源于它特有的位于体腔顶部的微管复合体。你可以在如下网址获得关于 Apicomplexa 的更多信息：http://ucmp.berkeley.edu/protesta/apicomplexa.html 或者 http://www.tulane.edu/~wiser/protozoology/notes/api.html。

疟原虫的全基因组序列已经被一个国际性组织所报道（Gardner 等，2002）。这项测序工作格外困难，因为该基因组内的 80.6% 是 AT，是所有的真核生物基因组中最高的。在基因之间的区域和内含子中，某些时候 AT 的含量达到了 90%。一种全染色体（而不是全基因组）鸟枪法测序技术被应用到测序工作中。即染色体首先被凝胶分离开来，DNA 被提取出来，包括了 1～3kb 长度的 DNA 鸟枪法文库。基因组大小是 22.8Mb，含有 14 个大小从 0.6～3.3Mb 的染色体。

Gardner 等（2002）鉴定了疟原虫中的 5268 个编码蛋白质的基因。这个数目同 *Schzos-*

accharomyces pombe 酵母的基因预测数目是一致的，虽然它的基因组要大了一倍。总的来看，大约每4300bp中就有一个基因。Gene Ontology Consortium 中大约含有40%的基因产物的定义（大约2100个）。然而剩余的60%的预测出来的蛋白质在其他真核生物中没有找到同源蛋白。这些蛋白质是潜在的药物作用靶点。例如，有些蛋白对于形成 apicoplast 是必要的。apicoplast 是一个质体的遗留，与叶绿体具有同源性，在类异戊二烯和脂肪酸的代谢中起作用。

在网上有许多研究疟原虫的资源（表16-14）。PlasmoDB（Bahl等，2003）是基因组数据的资源中心。

表16-14 疟原虫基因组网络资源

资　源	注　解	地　址
PlasmoDB	有关疟原虫的主要网页资源	http://www.plasmodb.org/
Links page	位于 NCBI	http://www.ncbi.nlm.nih.gov/projects/Malaria/related_links.html
P. falciparum Genome Project	位于 the Sanger Institute	http://www.sanger.ac.uk/Projects/P_falciparum/

与疟原虫基因组项目同时，另一个组织测定了啮齿类动物中的疟疾病原体——*Plasmodium yoellii yoelii*（Carlton等，2002）。这是一个重大的进步，因为疟原虫的生活世代不能脱离宿主，而 *P. yoelii* 可以。后者的基因组大小为23.1Mb，含有14个染色体。AT含量同样很高（77.4%）。基因组内编码的基因数目也差不多。当用BLAST将疟原虫的5268个预测蛋白跟 *P. yoelii* 的5878个蛋白进行比较时（E-value 阈值为 10^{-15}）发现了3310个直系同源蛋白。其中包括了能引起人类免疫系统反应的潜在疫苗抗原（Carlton等，2002）。

获取更多关于疟疾的信息请参考 http://www.wellcome.ac.uk/en/malaria 和 http://www.who.int/tdr/diseases/malaria/default.htm。

Charles Louis Alphonse Laveran 因为关于疟疾治疗的工作获得了1907年度的诺贝尔奖（http://www.nobel.se/medicine/laureates/1907/index.html）。更早些时候，Ronald Ross 因为关于疟疾的研究获得了1902年的诺贝尔奖（http://www.nobel.se/medicine/laureates/1907/index.html）。

疟原虫的基因组测序组织包括了以下单位：WellcomeTrust Sanger 研究所、基因组研究所、美国海军医学研究中心（NMRC，马里兰州）、斯坦福大学，一种黏菌（*Dictyostelium discoideum*）也含有很高的AT含量（如下文）。

NCBI 关于疟疾的基因和基因组网站位于 http://www.ncbi.gov/projects/Malaria/。

质体是一个有光合作用功能的细胞器。最著名的质体是普遍存在于绿色海藻和陆地植物中的叶绿体（Gilson 和 McFadden，2001）。见下文的植物部分。

类异戊二烯是一个5碳原子的化学分子，是许多自然化合物（包括类固醇、视黄醇和气味物质）的组成成分（RBp 和 OBP 都是能运输类异戊二烯的脂质运载蛋白）。

当疟原虫和 *P. yoelii* 的基因组序列都已知时，生物信息学和基因组学的方法如何应用到对这些生物的基础研究中呢？现在的数据中包括了几千个以前不知道的基因，给抗击疟疾提供了新的可能的途径（Hoffman等，2002）。

① 顶端质体是潜在的药物靶标。Zuegge等（2001）分析了定位到顶端质体的84个蛋白的末端氨基酸序列和102个与顶端质体无关的蛋白质（如细胞质内的分泌蛋白和线粒体蛋白）序列。他们使用主成分分析、神经网络和自组织网络（第7章）建立了顶端质体的定位信号的预测模型。

② 与其他亲缘关系较近的物种的比较基因组学研究获得了关于基因组结构、基因组组分以及其他基因组特性的重要信息。Carlton 等（2001）比较了来自疟原虫、*P. vivax* 和 *P. bergheid* 的 EST 和基因组相关序列（见第 2 章）。他们鉴定了其中表达量最高的一部分基因，得到了一些结果，如 *rif* 基因家族暗示了抗原的多样性。

③ *P. yoelii* 和 *P. falciparum* 之间的保守共线性分析覆盖了 16Mb 的序列，为这些病原体的进化提供了信息。Carlton（2002）用 MUMmer 程序（第 14 章）对蛋白质编码区域进行比对。保守共线性的图谱揭示了基因排列顺序的保守性，为分析染色体断裂点提供了可能，确认了某些基因的缺失（比如 *P. yoelii* 中的 *var* 和 *rif*）。

④ 人们可以研究那些引起抗原多样性和对免疫系统的逃避的基因。在 *P. vivax* 中，*vir* 基因有 1000 份拷贝，定位在亚端粒区域。*Plasmodium yoelii* 拥有 838 个 *yir* 的拷贝（Carlton 等，2002）。

⑤ 许多研究小组把蛋白质组技术用于分析疟原虫的四个生命阶段（孢子期、裂殖期、营养体期和配子期）的分析。Floren（2002）等鉴定了 2415 个蛋白质的表达，占到了所有预测基因的一半。另一个意料之外的发现是 *var* 基因和 *rif* 基因（过去被认为是与入侵免疫系统有关）在孢子体期也有大量的表达。综合起来说这些研究给出了不同阶段的蛋白质表达谱，为研究蛋白质的功能提供了信息。蛋白质组的研究也验证了在基因组 DNA 上进行的基因组学研究结果。Lasonder 等（2002）用质谱鉴定了一些蛋白质的序列，这些序列是在分析基因组 DNA 时预测的基因中所没有的。

⑥ 分析疟原虫的代谢通路作为治疗的靶点也是有可能的（Gardner 等，2002；Hoffman 等，2002）。所有研究过的物种都用异戊基二磷酸化酶合成类异戊二烯。一些植物或者细菌使用另外一种非典型的代谢途径，该途径利用了哺乳动物中所没有的 DOXP。Jomaa 等（1999）用 TBLASTN（用细菌的 DOXP 的蛋白质序列对疟原虫的基因组序列进行搜索）找到了一个疟原虫的直系同源基因。他们的研究表明该蛋白可能定位在顶端质体中，并且疟原虫对低浓度的该酶的抑制剂十分敏感。他们的研究还表明这些药物对于感染疟原虫的老鼠有抗疟疾作用。生物信息学方法在寻找更多的抗疟疾的药物的研究中将继续发挥很大的作用。

我们已经讨论了几种原生生物的基因组计划。更多的基因组计划正在进行中，如纤毛虫（Dessen 等，2001）和 alveolates 的四鞭毛虫类的基因组计划。正在测定许多原核生物基因组的基因组研究中心也在同时测定许多原生生物和其他真核生物的基因组。

> 引起小鼠疟疾症的疟原虫正在被测序，包括 *Plasmodiumberghei* 和 *Plasmodium chabaudi*（Hoffman 等，2002）。预测顶端质体的定位序列的数据库位于 http://gecco.org.chemie.uni-frankfurt.de/pats/pats-index.php。我们在第 5 章中讨论了 *vir* 基因（问题 2），我们用 BLAST 和 PSI-BLAST 对这个家族进行了评估。

> 草履虫基因组计划网址：http://paramecium.cgm.cnrs-gif.fr/。四鞭毛虫的测序计划网址：http://lifesci.ucsb.edu/~genome/Tetrahymena/。你可以在图 16-1 中看到这两种生物在进化树中所处的位置。T. gondii 数据库在 http://ToxoDB.org（Kissinger 等，2003）。
> TIGR 的真核生物基因组计划（http://www.tigr.org/tdb/euk/）包括了一系列原生生物和后生生物寄生虫，如疟原虫、*P. vivax*、*P. yoelii*、*T. gondii*、*T. brucei*、*T. cruzi* 和 *Schistosoma mansoni*。

16.3.5 植物基因组

16.3.5.1 概述

地球上有几十万种植物，分子进化树表明植物形成了真核生物的一个独立分支（见 viri-

diplantae，图 16-1)。所有的植物都是多细胞生物，因为它们是从由母体组织形成的具有多细胞结构的胚胎发育而来（Margulis 和 Schwartz，1998）。大部分植物都能进行光合作用，虽然也有少部分不能（如山毛榉，Epifagus）。

对植物基因组的分析能让我们了解到是什么样的分子基础让植物和动物有如此大的不同，如植物有特化的细胞壁、液泡、质体和细胞骨架。植物是固着生长的并且能进行光合作用。对植物基因组的测序可能会对这些基本的特性作出一些解释。

今天的植物的祖先是什么时候同动物、真菌和其他生物发生分支的呢？最早的生命证据出现在大概 38 亿年前（3.8BYA），而最早的真核生物的化石已经有 27 亿年的历史了。图 16-22 中的系统树描述了这些事件，Meyerowitz（2002）和 Wang 等（1999）的独立研究证明了上述事实。没有找到特别早期的植物化石，因此想要断定植物是什么时候分支出来的非常困难。许多研究者用基于蛋白质、DNA 或者 RNA 的序列的分子钟方法来推断分化时间。Wang 等（1999）使用了 75 个核内基因的序列来估计植物真菌和动物的分化时间。他们采用由找到的化石确定的哺乳动物和鸟类的分化时间（3.1 亿年前）来校准他们构建的进化树。他们发现动物与植物的分化时间为 15.47 亿年前，同动物与真菌分化的时间（15.38 亿年前）基本一致（图 16-22）。

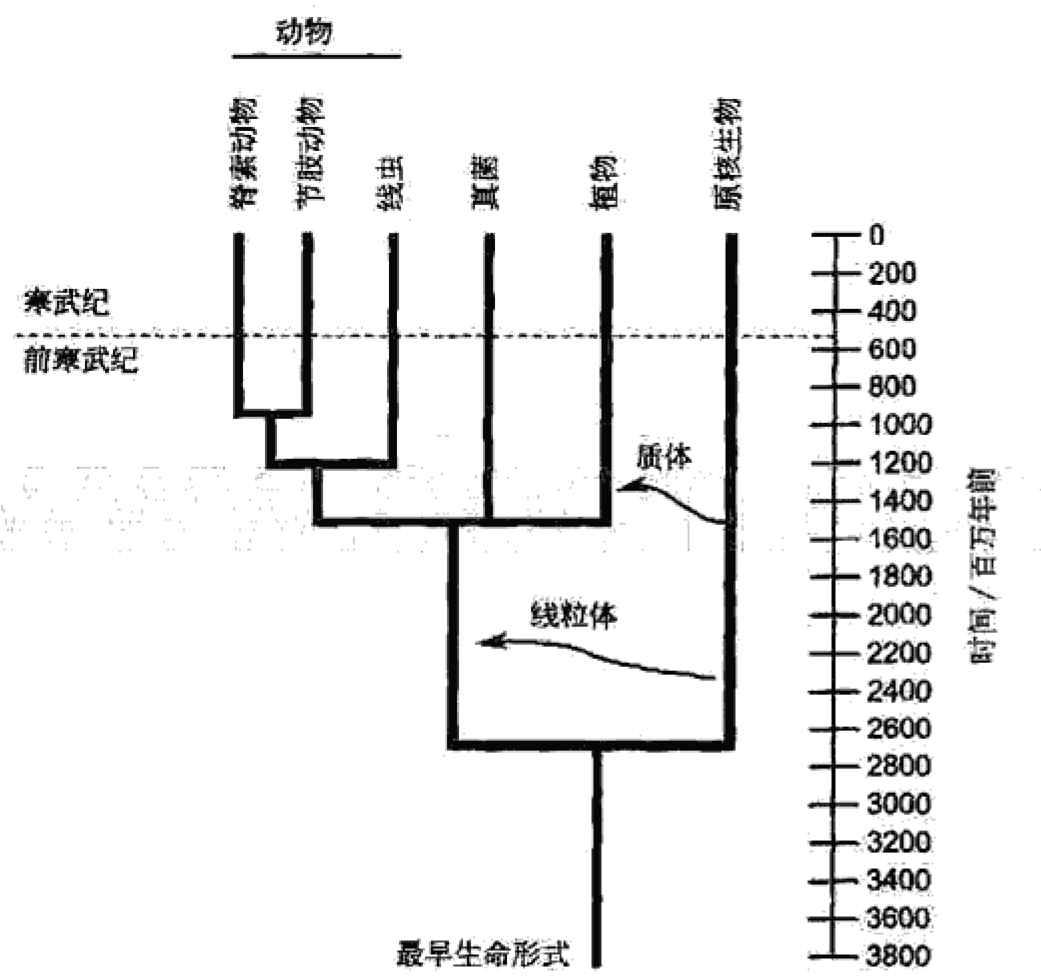


图 16-22 植物动物和真菌的进化。估计出来的植物动物和真菌的分化时间约为 15 亿年前（来自 Wang 等，1999；有修改）。在这个分化事件之前，一个单细胞的真核生物摄取了一个 α proteobacterium（现在的线粒体，在现有的植物、真菌和动物中都存在）。在大约 15 亿年前之后，植物的祖先获得了一个质体（叶绿体）。根据这个模型，后生生物的出现时间比根据化石所推断的时间要早 4 亿年。另外，蠕虫类动物（如线虫）比脊索动物（如脊椎动物）和节肢动物（如昆虫）先出现。资料来自 Meyerowitz（2002）和 Wang 等的独立研究，有修改，已获授权

Epifagus virginiana 叶绿体基因组已经被测序（NC_001568）（Wolfe 等，1992）。*Epifagus* 是山毛榉根部的一种寄生生物，其叶绿体基因组原有的光合作用功能已经丧失，它缺少其他植物的叶绿体基因组中所拥有的 6 个核糖体蛋白质和 13 个 tRNA 基因（Wolfe，1992）。
植物和动物在基因组成分上有显著的不同，如植物缺少胞间连丝以及编码胞间连丝的蛋白质（如角蛋白和 vimentin）的基因。

既然植物、动物和真菌是由一个共同的单细胞真核生物祖先分化而来，因此比较植物和动物的基因组就能使我们看到植物和动物是如何独立地进化成多细胞形式的（Meyerowitz，2002）。植物和动物的线粒体基因组是同源的，表明了它们的共同祖先细胞内已经含有了 proteobacterium（图 16-22）。在分化以后的另一个内吞事件中，蓝细菌侵入植物细胞形成了叶绿体。这个事件独立发生了多次（框 16-2）。同样地，试图给这些事件一个确定的时间是很困难的（Meyerowitz，2002）。许多动物门最早的化石出现的时间都是在大约 5.3 亿年前，这个现象被称为寒武纪大暴发。

框 16-2　令人惊奇的一种获取叶绿体的方式：吞食含有叶绿体的生物

叶绿体是植物体内一种质体（光合作用细胞器），它含有绿色的色素（叶绿素）。叶绿体可以把光转换成能量。一个关于叶绿体起源的主要假说是，在植物同动物和真菌分化后不久，一个真核细胞摄取了一个蓝细菌（图 16-22）。但是另一种机制也是常见的：一个真核细胞可以摄取一个已经含有叶绿体的藻类（Gilson 和 McFadden，2002）。这种被称为内共生的现象至少发生在 7 个不同的真核生物类别中：顶端复合体（上面已经讨论过）、chlorarachniophytes、隐形虫、腰鞭毛虫、眼虫藻、不等鞭毛属、hyptophytes（Gilson 和 McFaddan，2002）。许多含有叶绿体的植物和藻类都含有 3 个基因组：核基因组、叶绿体基因组和线粒体基因组。在隐形虫和 chlorarachniophytes 中还有第四个基因组：被吞入的藻类退化了的基因组。这第二个核被称为核子。

同那些细菌的基因组一样，核子的基因组也是小而高度简化的。Douglas 等（2001）测定了 *G. theta* 的核子基因组序列，仅有 551264bp。基因的密度非常高，平均每 977bp 就有 1 个基因。非编码序列区域非常短，只有一个假基因。有些基因（如编码 DNA 聚合酶的基因）已经丢失，而有些基因产物要进入质体，需要通过四层不同的膜。

G. theta 的环状质体 DNA 也是非常紧凑的。Douglas 和 Penny（1999）测定了这个长 121524bp 的基因组，发现 90% 的 DNA 都是编码 DNA，没有假基因和内含子。与之形成对比的是，水稻的质体基因组只有 68% 的区域是编码区。可以在 NCBI 网站上浏览 *G. theta* 的基因组（编号：NC000926），也可以将它同红藻的质体基因组进行比较（编号：NC000925）。这两个基因组显示出了高度的保守共线性。还可以把 *G. theta* 的质体基因组跟硅藻的质体基因组比较（编号：NC001713）。硅藻也是通过内吞获得的质体，但是它缺少核子。

已发现的最早的植物化石来自于志留纪（4.3 亿～4.8 亿年前）（Margulis 和 Schwartz，1998）。
被子植物是开花植物，种子由子房包被，子房发育成果实。单子叶植物是根据胚胎中只有一片子叶来定义的，代表性植物有水稻、小麦和燕麦。双子叶植物的胚胎有两片子叶，代表性植物有马铃薯和番茄。双子叶植物包括了花卉和树的大部分（除针叶树外）。被子植物的种系树请参见 http://www.mobot.org/MOBOT/Research/APweb/welcome.html，它包含了几十种进化树，并且可以浏览文字、植物图片和更多的参考资料。与被子植物不同的是裸子植物，它们的种子包被在球果中。

我们将在 NCBI 上对植物的分类开始我们的生物信息学和基因组学探索（表 16-15）。主要的植物类别之间的关系也可以通过用植物里的一个关键酶（核酮糖-1,5-二磷酸羟化酶）构

建的进化树来说明（图 16-23）。绿色植物的两大主要类别是 Chlorophyta（绿色藻类，如 *chlamydomonas*）和 streptophyta。Streptophyta 可以进一步分为苔藓类、地钱和被子植物（开花植物）。被子植物包括了单子叶植物和双子叶植物。

表 16-15 植物分类（Kingdom Viridiplantae）

类别	种类	门	目	举例
Chlorophyta (green algae)	Chlorophyceae	—	—	*Chlamydomonas reinhardtii*
	Pedinophyceae			
	Picocystophyceae			
	Prasinophyceae			
	Trebouxiophyceae			
	Ulvophyceae			
Streptophyta	Charophyta/Embryophyta	Charales	—	*Chara fragilis*
		Coleochaetales	—	*Coleochaete nitellarum*
		Embryophyta(plants)	Anthocerotophyta (hornworts)	*Megaceros aenigmaticus*
			Bryophyta(mosses)	*Physcomitrella patens*
			Marchantiophyta(liverworts)	*Asterella gracilis*
			Tracheophyta(vascular plants)	*Arabidopsis thaliana*
	Chlorokybophyceae			
	Klebsormidiophyceae			
	Mesostigmatophyceae			
	Zygnemophyceae			

注：1. 维管植物呈高亮状态。

2. 来自并改编于 NCBI 分类目录（http://www.ncbi.nlm.nih.gov/）。

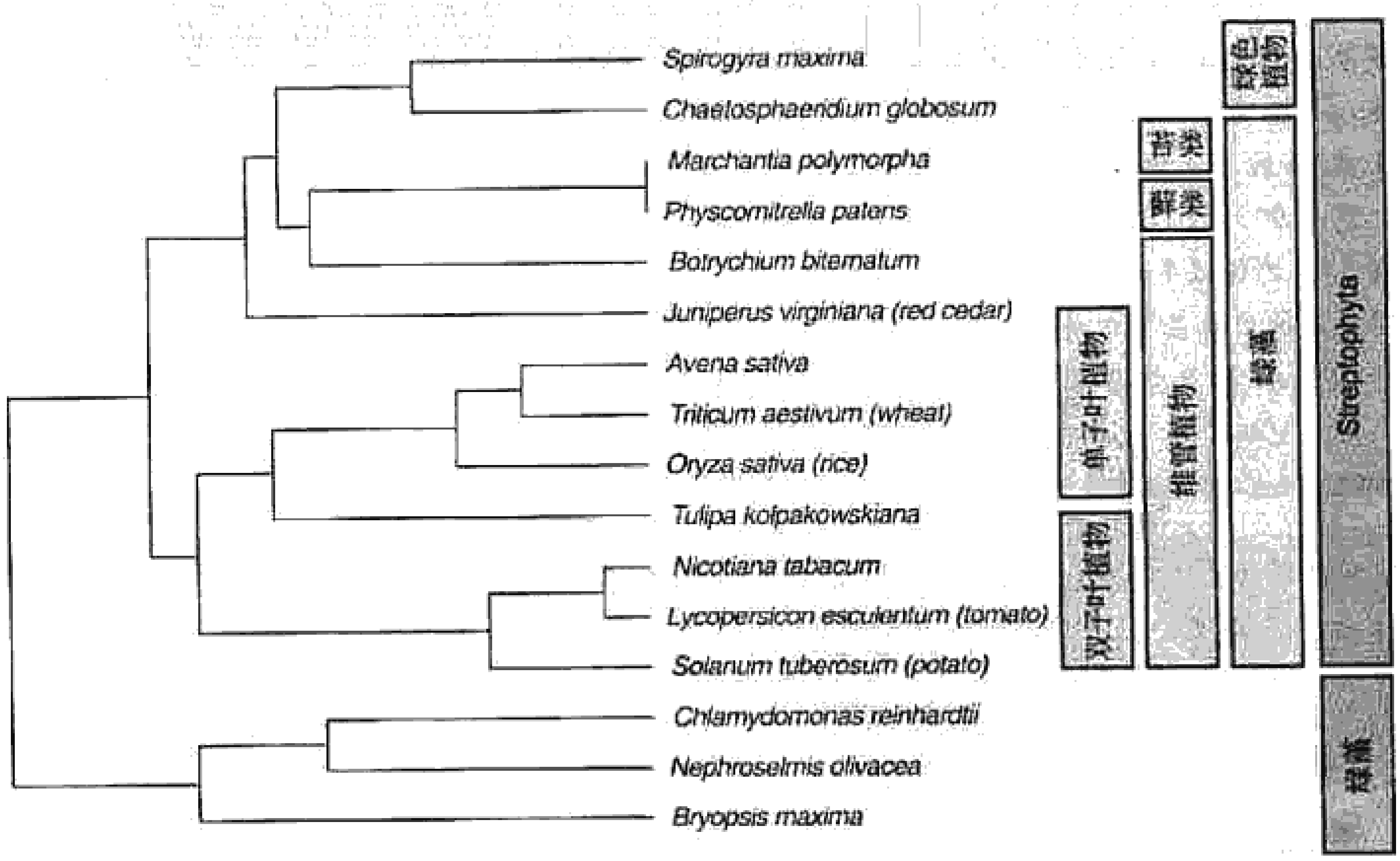


图 16-23 用 18SRNA 构建的进化树显示了动物和真菌的分支（图 16-1），与图 16-22 的结果是一致的

在开花植物中，拟南芥（双子叶植物）和两种水稻（单子叶植物）的基因组是最早被测序的两个植物基因组。另外一系列其他植物的基因组也正在测序当中（表16-16）。

16.3.5.2 拟南芥基因组

拟南芥是一种通常被认为是野草的水芹类植物，它作为基因组第一个被测序的植物而闻名。拟南芥已经被植物研究界作为一种模式生物来研究，因为它体积小、世代时间短、后代众多，并且便于基因水平上的操作。它是芥菜属的成员之一，芥菜属包括了山葵、椰菜、花椰菜和芜菁。它是250000种开花植物中的一种，大约在2亿年前分化出来（Walbot，2000）。比较基因组学研究将拟南芥的基因组同其他开花植物的基因组进行比较，以了解植物基因组的更多情况（Hall，2002）。

拟南芥的基因组大约长125Mb，与一些重要的农业作物［如小麦和大麦（16.5Gb和5Gb）］相比，它的基因组是非常小的（表16-16）。这使得它成为了第一个被测序的植物。拟南芥基因组项目组（2000）报道了几乎全部的基因组序列（115Mb）。拟南芥一共有5条染色体，包含了25498个预测的基因。拟南芥基因组的基因分布密度是大约每4.5kb有一个基因。

表16-16 主要植物测序项目

植　物	俗　名	基因组大小/Gb	与人类基因组大小比较
Arabidopsis thaliana	水芹	0.125	1/25.6
Avena sativa	燕麦	16	大5倍
Glycine max	大豆	1-2	约1/2
Hordeum vulgare	大麦	5	大1.7倍
Lycopersicon esculentum	番茄	1	1/3.2
Medicago truncatula	苜蓿	0.5	1/6.4
Oryza sativa	水稻	0.466	1/6.9
Triticum aestivum	小麦	16.5	大5倍
Zea mays	玉米	2.365	1/1.4

注：见NCBI植物资源（http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/PlantList.html）。许多植物基因组的大小与人类相当（3.2Gb）。

rubisco是指核酮糖-1,5-二磷酸羧化酶，它位于进行光合作用的植物叶绿体内，催化固碳反应的第一步。该酶催化核酮糖二磷酸和二氧化碳反应生成3-磷酸甘油分子。rubisco的基因名称是*rbcL*，一个典型的例子是水稻的该基因编码的蛋白（RefSeq编号：NP_039391）。
因特网上有许多模式植物的基因组数据资源，如MtDB（Lamblin等，2003）（http://www.medicago.org/）和玉米的数据库ZmDB（http://zmdb.iastate.edu）。更多的植物基因组数据库还包括Genoplante-Info（GPI）等（Samson等，2003）（http://genoplante-info.infobiogen.fr）。GrainGenes，一个关于小麦大麦、黑麦和燕麦的数据库，可见于http://www.graingenes.org（Matthews等，2003）。

根据TIGR和Crowe等（2003）对拟南芥基因组的重新注释，拟南芥中可能包含的基因数目略有增加。拟南芥含有的基因数目比果蝇（大约13000个基因）和线虫（大约19000个基因）的数目明显要多。这种现象可能跟植物内频繁发生的串联复制和片段复制事件有关。大概有11600个基因是核心部分，其他基因都是旁系同源基因（拟南芥基因组测序组织，2000）。

拟南芥的另一个令人惊讶的特性是它的基因组可能复制过2次。在基因组内有24个大的长度为100kb或者更长的复制过的片段，占了基因组总量的58%。将拟南芥的基因组同马铃薯的基因组进行比较发现，在4条染色体上有着相同的基因组成和顺序（Ku，2000）。

复制过的或者有3份拷贝的片段说明了整个基因组可能发生了二次或者更多次的大规模基因组复制。其中一次复制发生在远古而另一次则发生在大约1.12亿年前。当基因组复制发生后，随之而来的就是基因丢失。基因丢失事件降低了基因之间的共线性，给我们调查事件的发生经过和时间造成了困难（Simillion等，2002）。

双子叶植物（如拟南芥）同单子叶植物（如水稻）在大约2.0亿年前时分支。在双子叶植物中，紫菀苷和rosids大约在1.0～1.5亿年前分支。rosids包括了拟南芥、大豆和*M. truncatula*。紫菀苷包括了土豆。
这个测序工作是由TIGR、MIPS和TAIR（见表16-17）共同完成的。
TIGR数据库位于http://www.tigr.org/tdb/e2k1/ath1/，包括了许多资源，如注释数据库、片段复制图谱。TIGR给出了一个包括22485个基因的列表，以及关于另外16000个基因存在的不充分的证据。
拟南芥转录芯片数据（CATMA）数据库位于http://www.catma.org。

拟南芥的基因组数据资源在表16-17中列出。最全面的一个数据库是TAIR，包含了一个很大范围的服务（图16-24）（Rhee，2003）。这个站提供了对整条染色体的浏览，也提供了对单核苷酸多态性的描述（图16-25）。这个站点的格式被许多其他基因组计划所采用（框16-3）。

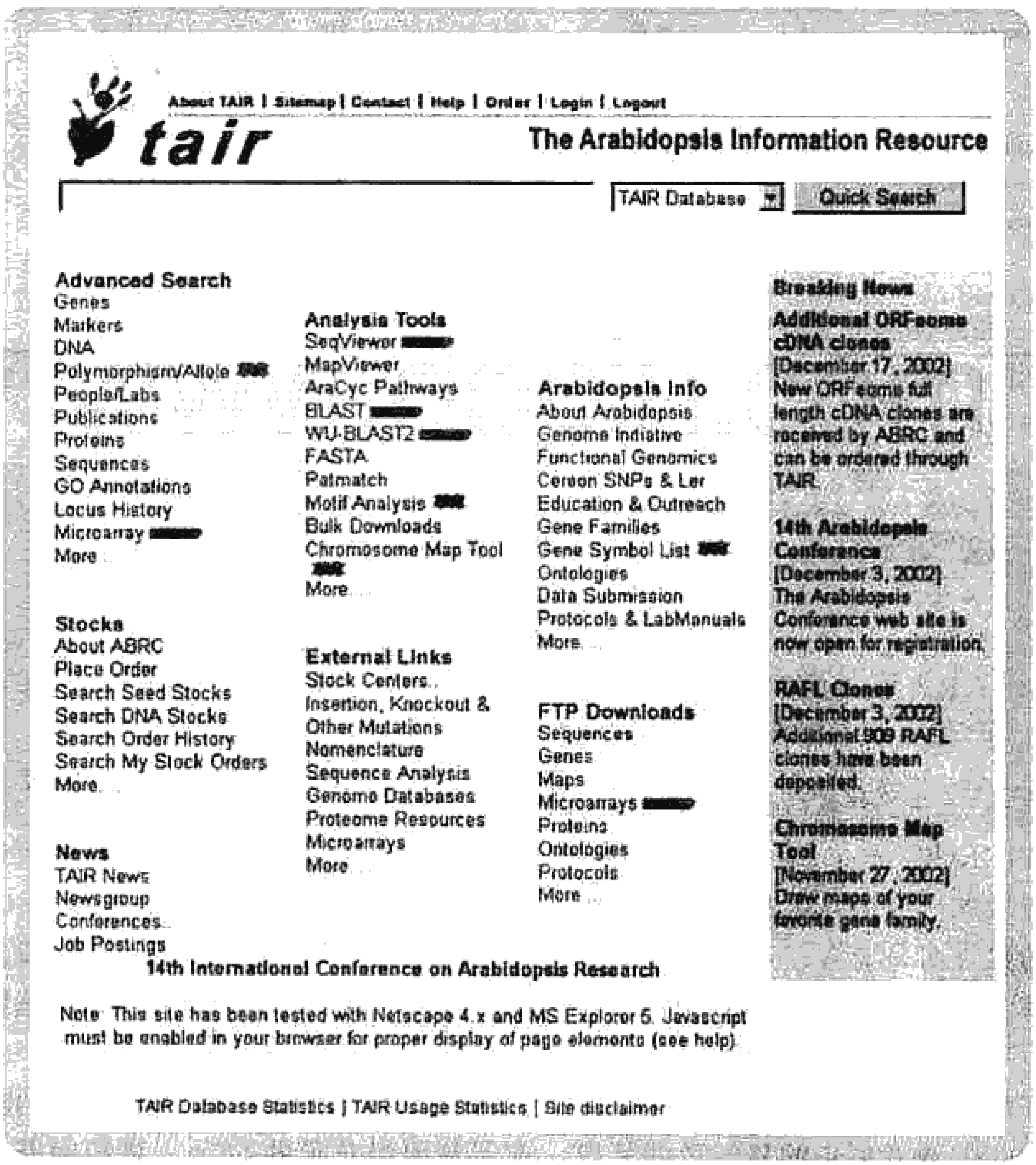


图 16-24 拟南芥信息资源库（TAIR）是一个关于拟南芥的全面的分子信息数据库

表 16-17 拟南芥基因组资源

资 源	注 解	地 址
TAIR	拟南芥资源	http://www.arabidopsis.org/
Arabidopsis thaliana Database	位于 TIGR	http://www.tigr.org/tdb/e2k1/ath1/
Arabidopsis thaliana Project	位于 MIPS	http://www.mips.biochem.mpg.de/proj/thal/proj/thal_overview.html
Arabidopsis genome analysis	位于 Cold Spring Harbor	http://nucleus.cshl.org/protarab/
SeedGenes	必要基因	http://www.seedgenes.org

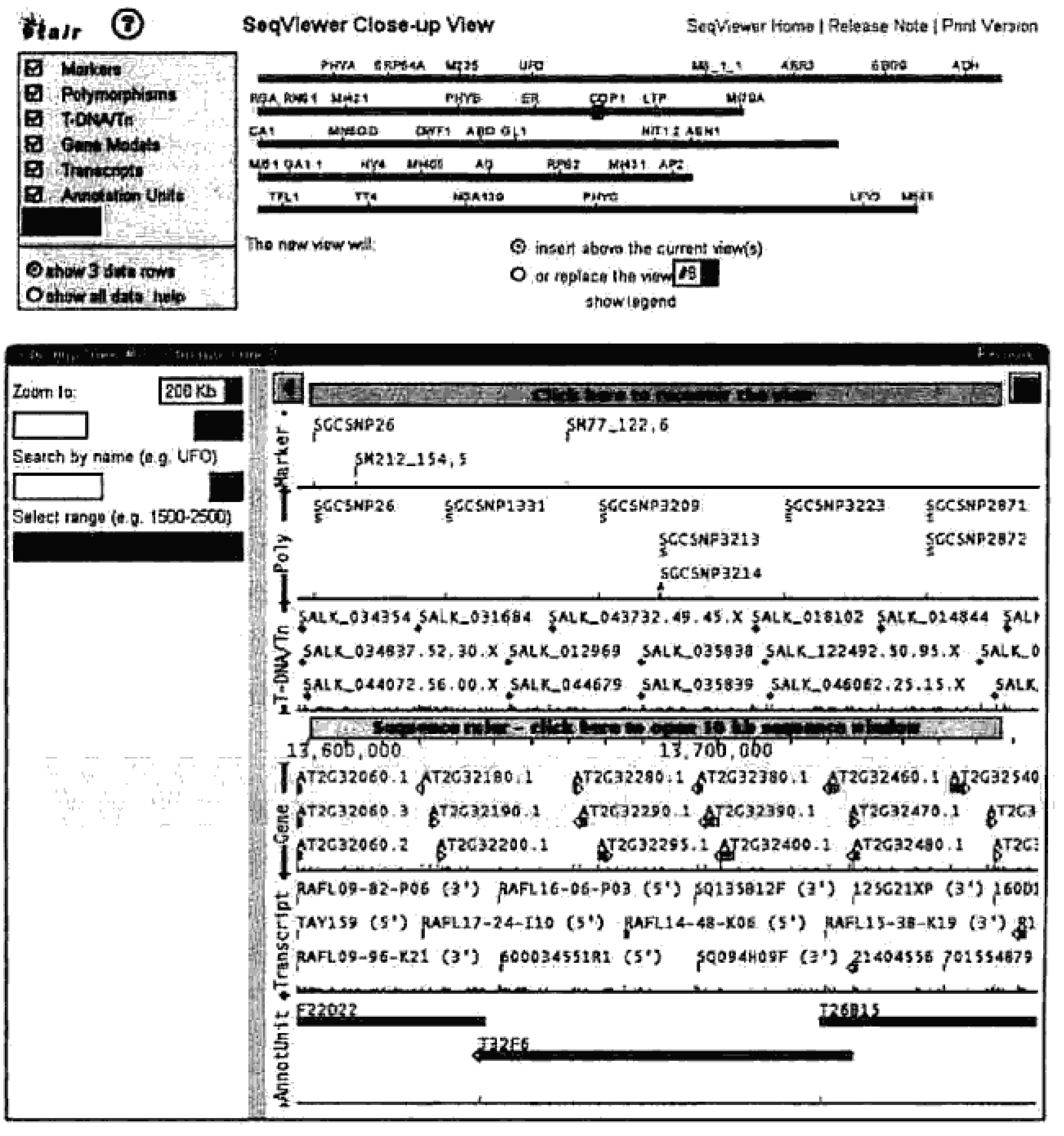


图 16-25 TAIR 网页浏览器提供了一个序列浏览器，既允许在基因组的大尺度上观看也可允许详细观察每个细节

其他的数据库（包括 SeedGenes）描述了拟南芥的必要基因（Tzafrir，2003）。

拟南芥中的大段片段复制是用 MUMmer（见第 14 章）和 tblastx（第 4 章）鉴定出来的。
禾本科包括了水稻、小麦、玉米、高粱、大麦、甘蔗、粟米、燕麦和黑麦，它总共有超过 10000 种的物种。

16.3.5.3 水稻

水稻曾经是地球一半以上人口的主食。水稻基因组是第二个被测序的植物基因组。它的基因组大小约为 430Mb，只有人类基因组大小的 1/8。尽管如此，它仍然是禾本科植物中基

框 16-3 真核基因组数据库

主要的拟南芥数据库（TAIR）使用了一个被许多测序计划公用的数据库模板（表 16-18）。我们已经在第 8 章中讨论了 EcoCyc，在第 15 章中讨论了酵母数据库 SGD。这些数据库既提供了在细节上的观察工具，也允许用户在整个基因组的大尺度上浏览数据（Gelbart，1998）。基因组统一图式（GUS）是另一种被广泛采用的模式（表 16-19）。许多数据库使用分布式注释系统（DAS），它允许一台计算机服务器从其他外部的服务器上获取病症和各种基因组数据。DAS 是由 Lincoln Stein 和 Robin Dowell 编写的，读者可以在 bioda. org（http://www. biodas. org）找到相关信息。它被 wormbase、flybase、ensembl 和 TIGR 广泛使用。

因组最小的植物。水稻也被当作单子叶植物的模式生物进行研究。

4 个小组得到了水稻基因组的草图（buell，2003），其中包括 2 个亚种的基因组草图。北京基因组中心领导的一个组织报道了水稻基因组的草图（*O. sativa L. ssp. indica*）（Yu，2002）。而另外一个组织报道了另一个不同亚种的水稻基因组序列（*O. sativa L. ssp. japonica*）（Goff，2002）。正如上面我们所讨论的，对已经完成的染色体 1 和染色体 4 的序列上的基因和其他特性的注释远比其他染色体的序列草图的注释更为详细充分。

许多数据库体提供了水稻的基因组和其他方面的数据，如 TIGR 和 MOsDB。这些站点也提供详细的基因组注释。

水稻基因组（印度亚属）的 GC 含量呈现出一个不同寻常的梯度。平均的 GC 含量是 43.3%，比拟南芥（34.9%）和人类（41.1%）稍微高一点。以 500bp 长度的序列的个数为 *x* 轴，以 GC 百分含量作为 *y* 轴作图，我们可以发现许多 GC 含量丰富的序列。这些富含 GC 的区域出现在外显子而不是内含子中，几乎在水稻基因组的每个基因中都可以发现一个 GC 含量特别丰富的外显子。每个基因的 5′端的 GC 含量通常要比 3′端的 GC 含量高出 25%。这些水稻基因组所特有的性质给使用从头预测基因的软件带来了主要的困难。

谷类是禾本科开花植物的种子，人们种植它们以获取它们的种子作为粮食。禾本科是单子叶植物，其形态为从小草到大树，生长期限可以是从每年生到多年生。参看 http://www. riceweb. org/glossary/Terms. html。
国际水稻基因组测序计划（IRGSP）测定了水稻日本亚属的草图 http://genome. sinica. edu. tw。另外有两个公司也在测定该基因组的序列，即孟山都公司（http://www. rice-research. org）和 Syngenta 公司（http://www. tmri. org）。北京基因组研究所领导测定了印度亚属水稻的草图（http://btn. genomics. org. cn/rice）。日本水稻基因组计划的网址为 http://rgp. dna. affrc. go. jp。美国水稻基因组测序计划的网址为 http://www. usricegenome. org/。
TIGR 水稻基因组数据库为 http://www. tigr. org/tdb/e2k1/osa1/。MIPS（*O. sativa*）数据库（MOsDB）为 http://mips. gsf. de/proj/rice。

表 16-18 使用模板的模式生物项目（GMOD）数据库（http://www. gmod. org/）

数据库	注解	地址
EcoCyc	大肠杆菌基因和代谢信息	http://EcoCyc. org/
FlyBase	果蝇网站	http://www. flybase. org/
Mouse Genome Informatics	小鼠主要资源	http://www. informatics. jax. org/
Rat Genome Database(RGD)	大鼠资源	http://rgd. mcw. edu/
SGD	见第 15 章	http://genome-www. stanford. edu/Saccharomyces/
TAIR	拟南芥信息资源	http://www. arabidopsis. org/
Wormbase	线虫网站	http://www. wormbase. org/

表 16-19 基因组统一模板平台（http://www.gusdb.org/）

数据库	注解	地址
AllGenes	人类和小鼠基因索引	http://www.allgenes.org/
EPConDB	胰脏项目计划	http://www.cbil.upenn.edu/EPConDB/
GeneDB	*S. pombe*、*Leishmania major* 和 *T. brucei* 的注释数据库	http://www.genedb.org/
PlasmoDB	疟原虫数据库	http://plasmodb.org/
RAD	RNA 数据库	http://www.cbil.upenn.edu/RAD2/

注：该平台用于一些物种数据库。

16.3.6 后生生物的最底层：黏菌

我们熟悉后生生物中的动物，包括蠕虫、昆虫、鱼类和哺乳动物。黏菌是一种社会性的阿米巴虫，作为一种真核后生动物它引起了人们很大的兴趣。参见图 16-1 和图 16-26，后者是利用 18S 核糖体 RNA 构建的进化树。

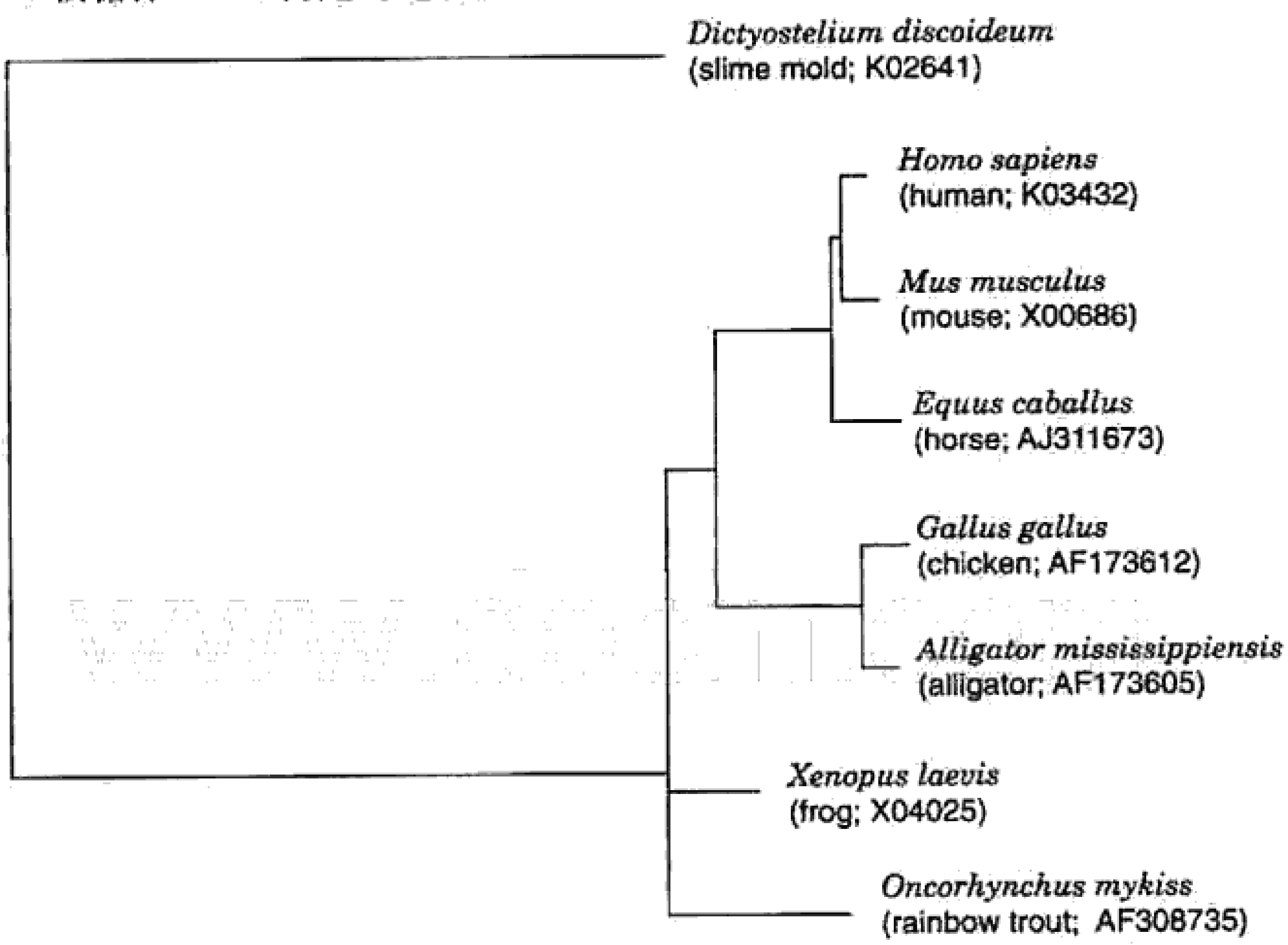


图 16-26 根据从 18S RNA 构建的进化树来看，黏菌与后生生物亲缘关系很近，这个结果跟图 16-1 中的结果很相近

生物学家研究黏菌是因为它引人注目的生活方式。在平时的条件下它是一种生活在土壤中的单细胞生物，当环境中的食物不够而导致饥饿发生时，它以类似脉冲的方式产生 cAMP，把大量的阿米巴虫聚集到一起，形成了一个具有其他多细胞真核生物特性的形态。它具有不同的细胞分化形式，对光和热有反应，也有发育的过程。

黏菌的基因组大小是 34Mb，分成 6 个染色体。其中有大约 100 份 90kb 长的回文序列的拷贝，编码了 rRNA 基因和线粒体基因组（55kb）。黏菌最大的染色体是 2 号染色体，由 8Mb 长的序列组成，已经被测序（Glockner，2002）。因为黏菌基因组中含有的 AT 含量与疟原虫一样，将近 80%，并且含有许多重复 DNA 序列，因此测序时采用了一种全染色体鸟枪法。染色体 2 含有 2799 个预测的蛋白，据此可以推断整个基因组大概有 11000 个基因。

基因的密度很高（约每2.6kb的基因组DNA中就有一个基因），内含子也较少（每个基因只有1.2个内含子）。这些内含子的AT含量为87%，而外显子中AT含量为72%。这个差异可能是由于内含子被切除引起的（Glockner，2002）。

请参见表16-20以找到更多的关于黏菌的网址。

表16-20 *Dictyostelium discoideum* 基因组资源

资源	注解	网址
Dictybase	*Dictyostelium* 和相关生物的资源信息中心	http://dictybase.org/
The *Dictyostelium* Genome Sequencing Project	Baylor College of Medicine	http://dictygenome.bcm.tmc.edu/
	Institute of Biochemistry I, Cologne	http://www.uni-koeln.de/dictyostelium/
	Department of Genome Analysis, IMB Jena	http://genome.imb-jena.de/dictyostelium/
	Sanger Institute	http://www.sanger.ac.uk/Projects/D_discoideum/
Dictyostelium links	NCBI	http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map00?taxid=44689

参见William F Loomis在http://dictybase.org/genomeseq.htm/的一篇文章以获取关于黏菌基因组测序计划的更多描述。
关于不同的分类系统，请参见Cavalier-Smith（1998）和Margulis和Schwartz（1998）。
关于根据细胞色素c氧化酶I构建的动物进化树请参见Hebert（2003）。Karl Leuckart（1822～1898）第一次将后生生物划分为6个门。

16.3.7 后生生物简介

后生生物包括了我们所熟悉的大部分动物。NCBI分类学网址把后生生物分为五个超门（表16-21），口侧动物随之被分成体腔动物（包括了人类和昆虫）、无体腔动物、假体腔动物（包括线虫）（表16-22）。可以在图16-22中查看线虫、节肢动物（昆虫）和脊索动物的关系。

体腔动物随之被分成了两大类：①原肢类动物，包括了节肢动物（昆虫）、环节动物和软体动物。我们将探讨最早被全基因组测序的原肢类动物，即昆虫黑腹果蝇和冈比亚疟蚊。②后口动物，包括了棘皮动物（如海胆）、半索动物（如蠕虫）和脊索动物（脊椎动物以及头索动物和尾索动物）。3个后口动物门有一个共同的祖先，大约在5.5亿年前寒武纪大暴发时发生了分化。我们将讨论脊索动物底部的一种动物——尾索动物海鞘，然后将讨论鱼类和老鼠两种脊椎动物的基因组。

表16-21 NCBI的后生动物界分类（http://www.ncbi.nlm.nih.gov/Taxonomy/）

超门	门	物种举例
Eumetazoa	**Bilateria**	*Homo sapiens*
	Cnidaria（珊瑚，水母）	*Polyorchis penicillatus*
	Ctenophora	*Beroe ovata*
Mesozoa	Orthonectida	*Rhophalura ophiocomae*
	Rhombozoa	*Dicyema misakiense*
Myxozoa	Myxosporea	*Myxobolus cerebralis*
Placozoa	Trichoplax	*Trichoplax adhaerens*
Porifera（sponges）	Calcarea	*Sycon raphanus*
	Demospongiae	*Geodia cydonium*
	Hexactinellida	*Aphrocallistes vastus*

表 16-22 口侧动物门

超门	门	物种举例
Acoelomata	Platyhelminthes(flatworms)	*Schistosoma mansoni*
Coelomata	***Deuterostomia*** ***(e. g. chordates, echinoderms, arrow worms)***	***Homo sapiens***
	Protostomia ***[e. g. mollusks, leeches, arthropods(insects)]***	***Drosophila melanogaster***
Pseudocoelomata	Acanthocephala(thorny-headed worms)	*Moniliformis moniliformis*
	Cycliophora	*Symbion pandora*
	Gastrotricha	*Draculiciteria tesselata*
	Kinorhyncha	*Pycnophyes kielensis*
	Nematoda(roundworms)	***Caenorhabditis elegans***
	Nematomorpha(horsehair worms)	*Gordius aquaticus*
	Rotifera	*Adineta vaga*

注：1. 最先被测序的动物为线虫、果蝇和人。

2. 改编自 NCBI 分类（http://www.ncbi.nlm.nih.gov/Taxonomy/）。

16.3.8 简单动物的分析：秀丽隐杆线虫

线虫是一种生活在土壤中的蠕虫，它被用来作为一种模式生物，因为它体积小（大约长 1mm）、容易繁殖（它的生活周期只有 3d），有被详细描述的不变的种系，便于遗传学操作。另外，它具有高等后生生物（如脊椎动物）所具有的各种生理特性，包括高级的中枢神经系统。许多线虫是寄生的，对线虫的了解可以为许多人类疾病的治疗提供信息。

> 口侧动物是双侧对称的动物。体腔是动物身体的空腔。无体腔动物缺少体腔。体腔动物有由中胚层形成的体腔，而假体腔动物有一个由其他组织形成的体腔。
>
> 有刺动物门是处于口侧动物外缘的一个门，大约在 6.5 亿～7.5 亿年前分化。它的成员包括了海葵、水螅、珊瑚、水母。CniBase 收集了关于有刺动物的基因组信息和资源信息，见 http://cnidbase.bu.edu/（Ryan 和 Finnerty，2003）。
>
> 雌雄同体的线虫的成体包括了 959 个细胞，其中包括了中枢神经系统的 302 个细胞。

研究线虫的另外一个优势是它的基因组非常小（只有 97Mb）。它是第一个被全基因组测序的动物，也是第一个被全基因组测序的多细胞生物（线虫基因组测序组织，CESC，1998）。基因组测序工作建立在已知的 5 个常染色体和 1 个 X 染色体的物理图谱的基础上。GC 含量是 36%，预测认为其基因组中有 19099 个编码蛋白质的基因，27%的基因组序列由外显子组成。大约 42%的蛋白质有非蠕虫动物的同源蛋白，只有 34%与其他的蠕虫蛋白同源。

线虫的蛋白质组包含了许多预测出来的含有 7 次跨膜结构域（7TM）的受体蛋白，有化学分子受体也有视紫红质受体（表 16-23）。这说明了在基因复制后，可能会产生有新的功能的蛋白质（Sonnhammer，Durbin，1997）。值得注意的是，许多线虫的蛋白在非后生生物（如植物和真菌）中不存在。

线虫的主要网络资源数据库是 WormBase，它是一个全面的关于线虫的数据库（Harris，2003）。WormBase 含有许多种类型的数据，包括：

① 基因组序列数据；

② 发育过程的种系；

③ 神经系统的连接方式；

表 16-23 线虫中十种最常见的蛋白结构域 Pfam 分类

InterPro	匹配的蛋白数目					名 称
	蠕虫①	植物②	苍蝇③	人类	酵母④	
IPR000168	638	0	0	0	0	线虫 7 跨膜螺旋化学物质受体(可能与嗅觉有关)
IPR000719	512	1046	265	688	118	蛋白激酶
IPR000276	403	6	86	826	0	类视紫质 GPCR 超家族
IPR000379	136	209	139	112	37	酯酶/脂酶/硫酯酶,活性位点
IPR0007114	127	83	114	100	67	主要协助运输蛋白超家族
IPR000387	106	25	37	139	10	酪氨酸特异性蛋白磷酸酶和双特性蛋白磷酸酶
IPR000324	104	0	17	18	0	维生素 D 受体
IPR005821	100	30	58	149	3	铁离子运输蛋白
IPR005828	96	90	103	75	53	通过底物运输蛋白
IPR006201	95	0	20	43	0	神经递质门控的铁离子通道

① 线虫。② 拟南芥。③ 果蝇。④ 酵母。

注：1. GPCR 表示 G 蛋白耦联受体。

2. 改变自 European Bioinformatics Institute (http://www.ebi.ac.uk/proteome/)。

④ 突变株表型，遗传标记，基因图谱数据；

⑤ 基因表达谱数据；

⑥ 文献资源。

2002 年的诺贝尔生理和医学奖授予了 Sydney Brenner、H. Robert Horvitz 和 John E Sulston，以表彰他们在使用线虫作为模式生物方面做出的开创性研究 (http://www.nobel.se/medicine/laureates/2002/)。大约有 300 种寄生性蠕虫能感染人体 (Cox, 2002)。目前已知的蠕虫有 20000 种，据估计大约有 100 万种蠕虫 (Blaxter, 1998, 2003)。

WormBase 网址：http://www.wormbase.org/。

线虫已经被用来进行许多功能基因组学的研究 (Fields, 1999; kim, 2001; Brook 和 Isaac, 2002)。6 个不同发育阶段的线虫的基因表达已经用基因芯片测出 (Hill, 2000)。大约有 10700 个开放阅读框 (56%) 已经被确认，这个数字同 EST 的数据是相吻合的。另外还有几千个在体内特定区域、不同发育阶段，以及不同生理条件下表达的基因。通过 RNA 干涉的方法，Kamath (2003) 阻止了 19000 多个基因的 86%的功能。他们在突变体表型 1 种鉴定出 1722 种基因。

16.3.9 第一个昆虫基因组：黑腹果蝇基因组

节肢动物也许是这个星球上最成功的一种真核生物，它们拥有的种类数量最多。它们包括有螯角的动物（如蝎子、蜘蛛、螨）和具颚昆虫（具有下颚的动物，如昆虫）(表 16-24)。果蝇成为一种重要的模式生物已经有一个世纪的历史了 (Rubin 和 Levwis, 2000)。果蝇是理想的研究对象，因为它的生活世代短（只有两周），具有复杂的表型（从眼睛颜色的变化到行为，发育和形态学上的变化），巨大的染色体在显微镜下能很容易地观察到。

表 16-24 节肢动物门 NCBI 分类 (http://www.ncbi.nlm.nih.gov/Taxonomy/)

超 门	种 类	超 门	种 类
Chelicerata	Arachnida(螨虫、蜱、蜘蛛) Merostomata(鲎) Pycnogonida(海蛛)	Mandibulata	Myriapoda(蜈蚣) **Pancrustacea**(甲壳类、昆虫类)

注：节肢动物是非脊椎原肢类动物（见表 16-22）Pancrustacea 分化为 Crustacea 和 Hexapoda。昆虫包括果蝇和冈比亚蚊。

已知的节肢动物共有 100 万种，但是根据估计，总数应该在 300 万～3000 万之间（Blaxter，2003）。
Thomas Hunt Morgan 因为对染色体在遗传中的作用的研究获得了 1933 年的诺贝尔奖，见 http://www.noble.se/medicine/laureates/1933。1995 年，Edward B Lewis、Christiane Nusslein-volhard 和 Eric F. Wieschaus 因为对早期胚胎发育中的基因控制的研究而分享了当年的诺贝尔奖。这些研究都是关于果蝇发育的（http://www.noble.se/medicine/laureates/1995/）。

果蝇的基因组序列测定使用了鸟枪法（Adams 等，2000）。在此之前，鸟枪法只是被用来测定小得多基因组，果蝇基因组测序的成功是一个突破性的进步。总长为 180Mb 的果蝇基因组被组织成 1 个 X 染色体（编号 1）、两个主要的常染色体（编号 2 和编号 3）、一个很小的第三个常染色体（编号 4，长度只有约 1Mb），以及一个 Y 染色体。大约全基因组的 1/3是异染色质（包括了简单重复序列、转座子元件和串联的 tRNA 基因）。这些异染色质分布在着丝点的附近以及整条 Y 染色体上。在常染色质和已染色质的交界处，包含了许多以前不知道的编码蛋白质的基因。

用 Genscan 对果蝇的全基因组进行预测发现了 17467 个基因，而用 Genie 则得到了 13189 个基因（Reese 等，2000；Adams 等，2000）。笔者们认为 Genscan 过高地估计了基因的数目。后来，基因组序列中去除了裂沟并且提高了质量。这导致产生了基因组序列的第 2 个和第 3 个版本。在第 3 个版本中，基因的数目（13676 个）同最初版本预测的数目基本没有差别（Misra 等，2002）。然而，估计的外含子数目有了明显的增加。另外有 85%的转录产物和 45%的蛋白质产物发生了变化。注释的进步的原因是更多的 EST 和全部的 cDNA 可以被比对到基因组上。这些研究分清了未翻译区域，而这项工作对于从头预测的方法来说是很困难的。

基因组注释的一项任务就是对编码蛋白质的基因的描述。另一项任务是描述这些基因的功能。这项任务可以通过找寻被预测基因的直系同源基因的方法来完成。Rubin 等（2000）用 blastp 对果蝇、线虫和酵母的蛋白质组进行了系统的搜索，得出了以下几个结论。

① 去除旁系同源后这些生物的蛋白质组的核心部分由无冗余的蛋白质组成。分别是 4383 个（酵母）、8065 个（果蝇）和 9453 个蛋白（蠕虫）。因此，虽然酵母是单细胞生物而果蝇和蠕虫都是多细胞生物，但其核心蛋白质组仅仅大了 1 倍。

② 近 30%的果蝇基因在线虫中有同源基因，20%的基因在酵母和线虫中有同源基因；这些基因编码的蛋白在所有的真核细胞中存在。

③ 果蝇的一半蛋白质在哺乳动物中有同源蛋白（E 值$<10^{-10}$），表明了果蝇同人的亲缘关系比同蠕虫更近（图 16-22）。

④ 果蝇的一部分蛋白质与酵母、线虫或哺乳动物的蛋白没有明显的相关性。

果蝇基因组的测定是多个单位的合作成果，包括了塞莱拉公司、伯克利果蝇基因组项目组（BDGP，http://www.fruitfly.org），以及欧洲果蝇基因组项目组（EDGP）（Adams，2000）。其中一部分工作是测定约 300 万条随机的约 500bp 长的基因组片段。
黑腹果蝇的主要数据库（以及其他种类的果蝇）Flybase：http://flybase.bio.indiana.edu/。
根据版本 3，13379 个编码蛋白的基因位于常染色质上，297 个位于异染色质上（Misra，2002）。
Rubin（2002）定义至少占全长 80%的序列有明显相似性的两个蛋白为直系同源蛋白，这个标准将低估实际的直系同源蛋白数目，因为许多匹配上的序列是很短的片段（图 8-3）。

⑤ 果蝇和线虫有 2200 个多结构域的蛋白质，而酵母中只有 672 个。类似的蛋白质组分

析将帮助发现不同物种间不同表型的分子基础。例如，果蝇和线虫有许多带有胞外结构域的蛋白质，它们在细胞间的通讯和细胞物质的交流中起作用，而酵母没有这种蛋白质。

16.3.10 第二个昆虫基因组：冈比亚疟蚊基因组

作为疟原虫的载体，冈比亚疟蚊非常有名。蚊子传播了许多种类的人类疾病，虽然大部分只是局限于热带地区（西尼罗河除外）（表 16-25）。

表 16-25 蚊子传播的人类疾病

疾病	蚊子种类	病例数
疟疾	*Anopheles gambiae*	5×10^8 人
登革热	*Aedes aegypti*	5×10^7 人/年
淋巴丝虫病	*Culex quinquefasciatus*, *Anopheles gambiae*	1.2×10^7 人
黄热病	*Aedes aegypti*	2×10^5 人/年
West Nile 病毒病	*Culex tarsalis*, *Culex pipiens*, other	约 3.4×10^3 人/年

注：来自 Budiansky（2002）和 Holt 等（2002）。西尼罗河病毒的数据来自美国 2002 年 1 月～12 月 30 日的病例数（疾病预防和控制中心）。

Holt（2002）报道了用全基因组测序法测定的冈比亚疟蚊的一个株系的基因组。该基因组大小为 278Mb，组成一个 X 染色体（编号 1）和两个常染色体（编号 2 和编号 3）。测序工作的一个困难是基因组中基因的高度可变性（这被称为单核苷酸差异）。因此两个丰度相当的单倍体可以形成一个嵌合体基因组结构。与之相反，黑腹果蝇和小鼠的基因都是完全同型结合的。

可以在 Ensembl 基因组浏览器上观看冈比亚疟蚊的基因组。Ensembl（第 12 章）和塞莱拉德的注释认为其中有 13683 个基因。同其他基因组一样，冈比亚疟蚊的基因组注释中也含有许多不完全和不准确的基因（Holt，2002）。

单元型是同一个染色体上一组基因座很相近的等位基因组合，它们趋向于共同遗传给子代。
Ensembl 基因组浏览器可以用来查看蚊子的基因组（http://www.ensemble.org/anopheles_g_gambiae/）。Ensembl 的主页（http://www.ensembl.org/）包括了人类、小鼠、大鼠、果蝇、线虫、斑马鱼和河豚的基因组。我们将在第 17 章讨论这一重要站点。
我们在第 6 章描述了果蝇 Down 综合征细胞黏附分子（DSCAM），它是一个通过可变剪切编码了多达 38000 个不同蛋白质的基因（NP_523649）。冈比亚疟蚊基因组中也有大量的可变剪切现象（Zdobnov，2002）。见 GenBank 蛋白质编号 EAA05472。
脊索动物中基因组最小的是 *Oikopleura dioica*，大约在 51～72Mb 之间（Seo，2001）。这种居住在海洋里的生物可能含有 15000 个基因。

冈比亚疟蚊的基因组大小为 278Mb，大约两倍于黑腹果蝇的基因组。这个大小上的差别是因为基因间的 DNA 序列，此外，与 Anopheles 种属和其他的果蝇种属相比，黑腹果蝇的基因组似乎经历过一次缩减（Holt 等，2002）。冈比亚疟蚊和黑腹果蝇大约在 2.5 亿年前分化（Zdobnov，2002）。这两个基因组中一半以上的基因都是直系同源的，氨基酸序列的总体相似性为 56%。为了做一个比较，发育成人和河豚的两个分支于 4.5 亿年前分化，但是这两个种类的蛋白质序列相似性为 61%。因此，昆虫的蛋白质分化比脊椎动物快。一个引人注意的问题是，冈比亚疟蚊吸食人血的特性以及鉴定药物靶点。因此鉴定节肢动物特有的和蚊子特有的基因是非常重要的（Zdobnov，2002）。

16.3.11 通向脊椎动物之路：海鞘

脊椎动物包括了鱼类、两栖类、爬行类、鸟类和哺乳类。所有这些生物都有分节的脊柱。脊椎动物是从哪里起源的呢？脊椎动物是脊索动物的一支。脊索动物体内都有脊索。海鞘是尾索动物，它是脊索动物的一种但是不是脊椎动物。海鞘是一种雌雄同体的非脊椎动物，它给我们提供了一个观察从非脊椎动物到脊椎动物的转变过程的观察窗口（Holland，2002）。

Dehal 等（2002）用全基因组鸟枪法得到了海鞘的基因组草图。海鞘基因组大小为 160Mb，大约是典型真菌生物基因组的 12 倍，但是只是人类基因组的 1/20。海鞘的 14 个染色体中含有 15852 个预测的基因，大部分预测出来的基因都有 ESTs 数据支持。

海鞘的基因组数据允许我们将原口动物和后口动物进行比较。将近 60% 的海鞘基因具有原口动物基因中的同源物。几百个基因具有非脊椎动物的同源基因但是没有脊椎动物的同源基因。2570 个海鞘基因（大约占总数的 1/6）有脊椎动物的直系同源基因但是没有原口动物的直系同源基因，这些基因是在后口动物种系分化成为脊椎动物、头索动物和后索动物之前出现的。有 3399 个海鞘基因（大约占总数的 1/5）既没有脊椎动物同源基因也没有非脊椎动物同源基因，因此可能是在后口动物分化之后出现的特有基因。

海鞘中含有各种基因，它们在一些生物过程（如凋亡、程序性细胞死亡）、甲状腺功能、神经功能和肌肉收缩功能中发挥作用。这给我们用比较基因组学研究脊索动物中那些重要的基本基因提供了一个机会。例如，神经元通过从突触前末端释放含有神经递质乙酰胆碱的突触小泡同肌肉联系。神经递质在突触中扩散，然后结合到突触后末梢的受体上。海鞘包含了关于神经递质的编码基因、合成乙酰胆碱的酶、乙酰胆碱转运酶、突触小泡的蛋白和神经递质受体。

C. intestinalis 的基因组主页在 http://genome.jgi-psf.org/海鞘 4/海鞘 4.home.html.。该基因组的 GenBank 编号是 AABS00000000，可以在 NCBI 上找到海鞘的 BLAST 服务页面。TIGR 有一个关于 *C. intestinalis* 的基因列表（http://www.tigr.org/tdb/tgi/cingi/）。Ghost 数据库包括了 *C. intestinalis* 的基因组数据和表达谱数据（http://ghost.zool.kyoto-u.ac.jp/indexr1.html）。

海鞘蛋白质（BAB85859）与人类的胆碱乙酰转化酶（NP065574）有 47% 的相似性。海鞘的一个基因（编号 AB071998）编码了一个与人类乙酰胆碱转运酶（NP003046）有 56% 相似性的蛋白。其他还有许多基因在非脊椎动物的神经传递中起作用。

16.3.12 鱼的基因组：河豚

日本河豚是一种脊椎动物，它的基因组非常紧凑。这种硬骨鱼有一个 365Mb 大小的基因组，只是人类基因组的 1/10（Aparicio，2002）。但是，河豚和人类在大约 4.5 亿年前分化，预测结果认为这两种物种有着差不多数目的基因。

认为河豚基因组很紧凑的原因有如下几点（Aparicio，2002）。

① Repeatmasker 的分析结果显示，散布性的重复序列只占了河豚基因组的 2.7%。而在人类基因组中，散布性的重复序列占了大约 45%。真核生物中所有已知的转座子元件的种类都可以在河豚基因组中找到。最常见的重复序列是类似于 LINEd 的 *Maui* 元件（6400 份拷贝），而人类基因组中最常见的 *Alu* 序列有多于 100 万份的拷贝。

② 内含子相对较短。75%的河豚内含子长度不到425bp，而人类基因组中75%的内含子小于2609bp。在河豚中，大约500个内含子长度大于10kb，而人类基因组中超过12000个内含子大于10kb。

③ 编码基因的序列占了108Mb长度，为整个基因组DNA（320Mb）的1/3。这个比例远远高于人类和小鼠基因组中基因的比例。

河豚基因组国际测序组织负责河豚的测序工作，关于河豚基因组数据的主网站是 http://genome.jgi-psf.org/fugu6/fugu6.info.html。
河豚基因组浏览器位于 http://www.ensembl.org/Fugu_rubripes/。由欧洲生物信息学中心的 Wellcome Trust Sanger 研究所开发，它是河豚基因组和其他基因组的一个主要站点。
一个收集了超过27000种鱼类数据的数据库位于 http://www.fishbase.org/。
另外两个鱼类的基因组计划是关于斑马鱼和淡水鱼青鳉的。有关斑马鱼的信息请参考 http://zfin.org（Sprague，2003）和 NCBI（http://www.ncbi.nlm.nih.gov/genome/guide/D_rerio.html）。关于青鳉的基因组信息请参见 http://www.dsp.jst.go.jp/MGI 以及青鳉表达谱模式数据库（http://medaka.dsp.jst.go.jp/MEPD）（Henrich，2003）。

16.3.13 分析哺乳动物基因组：小鼠基因组

Michael Zhang 实验室的 Xuan（2002）将小鼠的两个版本的基因组数据进行了比较，并且开发了小鼠和人类基因组浏览器工具（http://143.48.7.130/cgi-bin/gbrowse?source=cse）。塞莱拉公司的测序工作包括了5个小鼠种系：A/J2、DBA2/J、129X1/Svj、129S1/SvlmJ 和 C57BL/6J（见 http://www.celera.com/）。小鼠基因组的 GenBank 的编号是 CAAA01000000。可以在第17章中的3个人类基因组主要站点获得它的信息：http://www.ensembl.org/Mus_musculus/、http://genome.ucsc.edu/，和 http://www.ncbi.nlm.nih.gov/。

小鼠的基因组的测序和分析是生物学史上的一个里程碑。继人类之后小鼠成为第二个全基因组被测定的哺乳动物。小鼠是用来理解人体奥秘的绝佳模式动物。

① 值得注意的是，这两个物种大约在0.75亿年前分化，在小鼠的30000个基因中只有300个在人类基因组中找不到同源基因。

② 除了成千上万个编码蛋白质的直系同源基因外，人类基因组和小鼠基因组还有着许多同源的非编码DNA序列。这些保守序列给我们了解调控区域和非编码基因提供了条件。（Hardison，1997；Dermitzakis，2002；Dehal，2001）。

③ 小鼠和人类有着许多相同的生理特性。因此小鼠是研究成千上万种人类疾病（从感染性疾病到复杂的功能障碍性疾病）的最佳生物。

④ 有超过1000种自然突变的不同的小鼠种系。我们也可以用随机突变的方法（如用诱变剂或者辐射诱变）来获得新的小鼠突变种系。通过一些直接的方法（如转基因、基因敲除和基因敲入等），突变和其他基因水平上的修饰也可以应用到小鼠上。

两个独立的小组分别测定了小鼠的全基因组序列：小鼠基因组测序组织（Waterston，2002）和塞莱拉公司（未发表）。这些版本的数据由 Xuan（2002）进行了比较。他们选取了8300个拥有 RefSeq 编号的小鼠条目作为查询序列，并使用 BLAT 比较了两个版本数据的覆盖率和准确率。大部分 mRNA 都能在两个版本的数据中很好地匹配上。

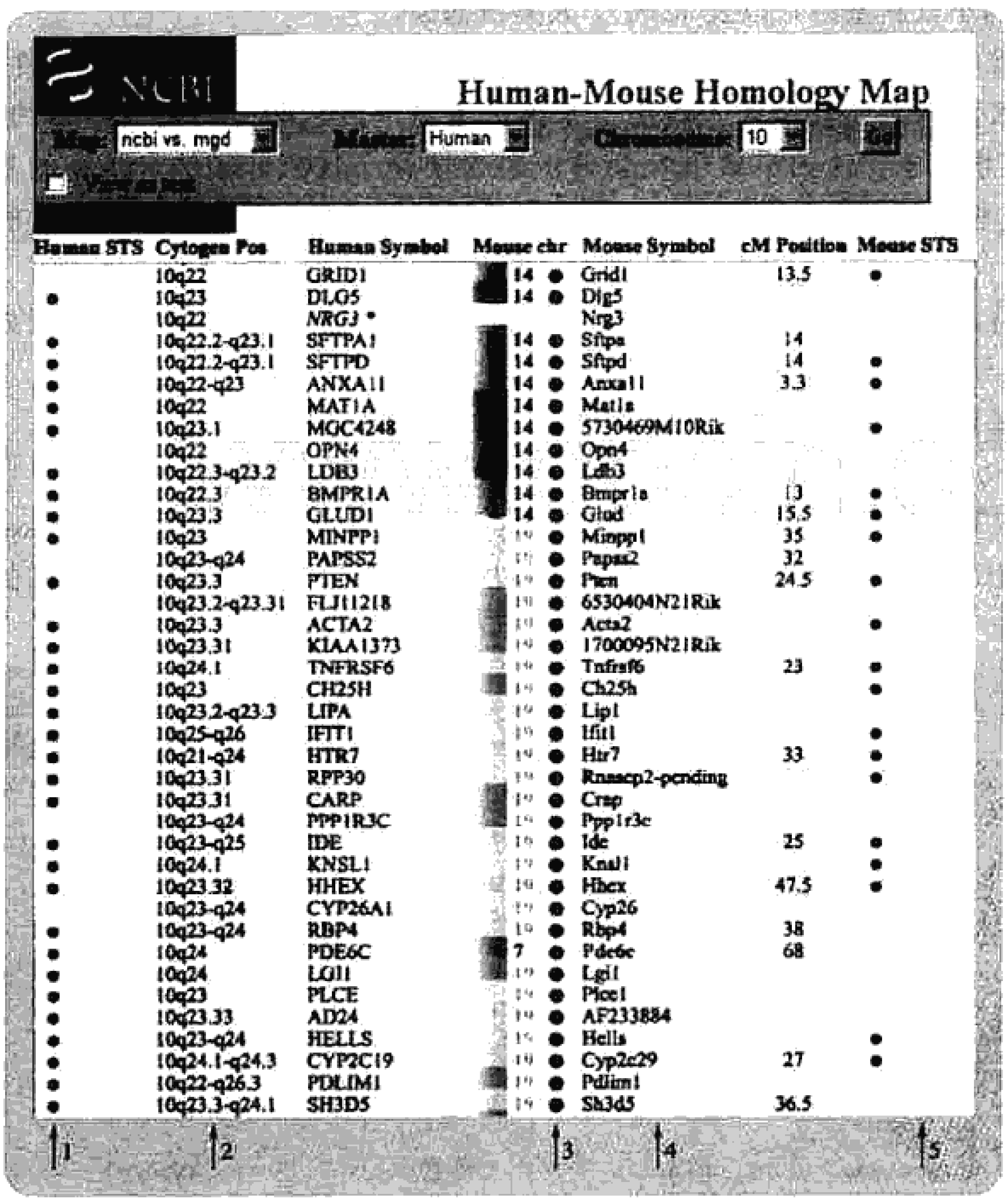
我们将讨论公共的测序组织的数据，因为它们是免费和公开的。Waterston 等（2002）测定了 B6 种系的一只母鼠的基因组。

他们的测序方法包括了全基因组鸟枪法（七层覆盖）和层次性鸟枪法（通过测定定位到染色体上的 BAC 克隆的序列）。综合数据涵盖小鼠基因组的大部分。在 RefSeq 的 cDNA 数据中，99.3%的 cDNA 可以同基因组数据比对起来。另外，Waterston（2002）还把另一个独立测出的 16 号染色体的数据也加入到 Refseq 数据库中（Mural 等，2002）。

Waterston 等（2002）提出了小鼠基因组测序工作的 11 个主要结论。

① 小鼠的常染色质长度为 2.5Gb，比人类的常染色质要小 14%（2.9Gb）。同本章内讨论的其他物种的紧凑的基因组相比，小鼠基因组（与人类基因组一样）中大约每 100000bp DNA 序列中有一个基因。GC 含量差不多，平均值为 42%（小鼠）和 41%（人类）。一共有 15500 个 CpG 岛，其中半数以上也能在人类基因组中找到（见 17 章）。

② 超过 90%的人类和小鼠基因组序列是保守共线性的。在小鼠和人类于 0.75 亿年前分化后，染色体 DNA 在每个物种中重排。然而，大段的 DNA 序列之间仍然存在相关性。作为一个可视的例子，NCBI 提供了一个人类/小鼠基因组同源性浏览器。选择人类基因组 10，将会看见一个人类基因的图谱（包括了视黄醇结合蛋白）（图 16-27），其中的基因与小鼠的 14 号、19 号和 7 号染色体上的基因同源。



Human STS	Cytogen Pos	Human Symbol	Mouse chr		Mouse Symbol	cM Position	Mouse STS
	10q22	GRID1	14	●	Grid1	13.5	●
●	10q23	DLG5	14	●	Dlg5		
	10q22	*NRG3* *			Nrg3		
●	10q22.2-q23.1	SFTPA1	14	●	Sftpa	14	
●	10q22.2-q23.1	SFTPD	14	●	Sftpd	14	●
●	10q22-q23	ANXA11	14	●	Anxa11	3.3	●
●	10q22	MAT1A	14	●	Mat1a		
●	10q23.1	MGC4248	14	●	5730469M10Rik		●
	10q22	OPN4	14	●	Opn4		
●	10q22.3-q23.2	LDB3	14	●	Ldb3		
●	10q22.3	BMPR1A	14	●	Bmpr1a	13	●
●	10q23.3	GLUD1	14	●	Glud	15.5	●
●	10q23	MINPP1	19	●	Minpp1	35	●
	10q23-q24	PAPSS2	19	●	Papss2	32	
●	10q23.3	PTEN	19	●	Pten	24.5	●
	10q23.2-q23.31	FLJ11218	19	●	6530404N21Rik		
●	10q23.3	ACTA2	19	●	Acta2		●
●	10q23.31	KIAA1373	19	●	1700095N21Rik		
●	10q24.1	TNFRSF6	19	●	Tnfrsf6	23	●
●	10q23	CH25H	19	●	Ch25h		●
●	10q23.2-q23.3	LIPA	19	●	Lip1		
●	10q25-q26	IFIT1	19	●	Ifit1		●
●	10q21-q24	HTR7	19	●	Htr7	33	●
●	10q23.31	RPP30	19	●	Rnasep2-pending		●
●	10q23.31	CARP	19	●	Crap		
	10q23-q24	PPP1R3C	19	●	Ppp1r3c		
●	10q23-q25	IDE	19	●	Ide	25	●
●	10q24.1	KNSL1	19	●	Knsl1		●
●	10q23.32	HHEX	19	●	Hhex	47.5	●
	10q23-q24	CYP26A1	19	●	Cyp26		
●	10q23-q24	RBP4	19	●	Rbp4	38	
●	10q24	PDE6C	7	●	Pde6c	68	
●	10q24	LGI1	19	●	Lgi1		
●	10q23	PLCE	19	●	Plce1		
●	10q23.33	AD24	19	●	AF233884		
●	10q23-q24	HELLS	19	●	Hells		●
●	10q24.1-q24.3	CYP2C19	19	●	Cyp2c29	27	●
●	10q22-q26.3	PDLIM1	19	●	Pdlim1		
●	10q23.3-q24.1	SH3D5	19	●	Sh3d5	36.5	

图 16-27 NCBI 的人类和小鼠的同源图谱展示了多个不同的保守共线性的图谱。人类和小鼠都可以用来作为主体，人类 10 号染色体是主体染色体（参考序列）。输出结果包括了人类 EST 数据（箭头 1），到 LocusLink 中人类数据的链接（箭头 2）和到小鼠数据的链接（箭头 4），到两个序列 BLAST2 比对结果的链接（箭头 3），以及到小鼠图谱浏览器的链接（箭头 5）

③ 40%的人类基因组序列可以在核苷酸水平上同小鼠的基因组匹配。其中展示了两个基因组共有的大部分同源基因。对于 12845 个同源基因来说，70.1%的氨基酸序列是一致的。

④ 通过比较基因组中成千上万的重复 DNA 序列与其远古祖先的对应序列的差异，我们可以估计每个基因组中的中性替代速率。人类的平均替代速率是每个位点 0.17 次，而小鼠的平均替代速率是每个位点平均 0.34 次。此外，小鼠的基因组中的短片段（小于 50bp）插入和删除的发生速率比人类高出 1 倍。

⑤ 处于选择压下的哺乳动物基因组中的小片段（50～100bp）大约占 5%。这个数字是用中性突变速率同基因组中的序列保守程序比较得出的。因为 5%这个数字大于编码蛋白质的基因所占的比例，因此基因组中那些非编码区域（如调控序列）也处于选择压之下。调控区域序列（如调控肝脏中的基因特异表达和肌肉中基因特异表达的调控序列）在小鼠和人类基因组中要比那些中性 DNA 区域保守得多，虽然要比那些编码蛋白质的区域保守性要低。

⑥ 哺乳动物基因组是以多种方式进化的，基因组中不同的序列其分化的速率也不同。中性的突变速率在各个染色体上各不相同，在 X 染色体上最低。GC 含量高的区域替代速率要相对较高。

⑦ 小鼠和人类的基因组都包含了大约 30000 个编码蛋白质的基因。80%的小鼠基因都含有一个人类同源基因。不到 1%的人类基因在小鼠中没有同源基因，反之亦然。测序工作发现了 9000 个以前未知的小鼠新基因，以及 1200 个人类新基因。

⑧ 大量基因家族扩增事件在小鼠基因组中发生，如气味分子受体基因家族扩增事件。小鼠基因组中 20%的该家族基因是假基因，表明了基因扩增和基因删除之间的互相制约。脂质运载蛋白也有基因扩增现象，比如小鼠 X 染色体上包含了一个气味分子结合蛋白的基因簇，而人类中没有。这些基因扩增也许可以部分解释灵长类动物和啮齿类动物在生理上的不同。

⑨ 哺乳动物的某些特定蛋白的进化速率非常快。例如，参与了免疫系统反应的基因处于正的选择压下可促使其快速进化。

⑩ 在人类和小鼠的基因组中找到了相似的重复性 DNA 序列。我们将在第 17 章讨论。

⑪ 公共组织报道了 80000 个单核酸多态性（SNPs）。我们将在第 18 章讨论。

万维网上最大的小鼠资源数据库是 MGI 数据库和它的相关站点（见侧栏）（Blake，2003）（图 16-28）。对 *RBP4* 的搜索可以得到一个对该基因非常详细的描述，包括了它们的图谱位置、性状和表达谱（图 16-29）。更多的关于小鼠的资源请参见表 16-26 和表 16-27。

选择母鼠进行测序保证了 X 染色体和常染色体的相同覆盖率。人类和小鼠的 Y 染色体都很小，包含了许多高度重复的 DNA 序列。

你可以从 NCBI 的网址 http://www.ncbi.nlm.nih.gov/Homology 上观看小鼠和人类的同源图谱。

小鼠基因组测序组织（Wateston，2002）定义了共线性区域为"两个物种中某一个染色体上一系列标记顺序相同的最大区域"。他们鉴定了 558000 个同源物和高度保守的序列标记，一共占到小鼠基因组的 7.5%。

Nadeau 和 taylor（1984）估计了小鼠和人类基因组中大概有 180 个保守共线性的区域。Wateston（2002）提供了 342 个这种区域的证据，每个区域的长度都大于 300kb。

小鼠基因组测序组织提供的小鼠基因组数据位于 http://www.ensembl.org/Mus_musculus/。塞莱拉公司的数据位于 http://www.celera.com。

MGI 位于 http://www.informatics.jax.org，由 Jackson 实验室负责维护（http://www.jax.org）。MGI 包括了小鼠基因组数据库（MGD）、基因表达数据库（GDX）、小数基因组测序（MGS）和小鼠肿瘤生物数据库（MTB）。

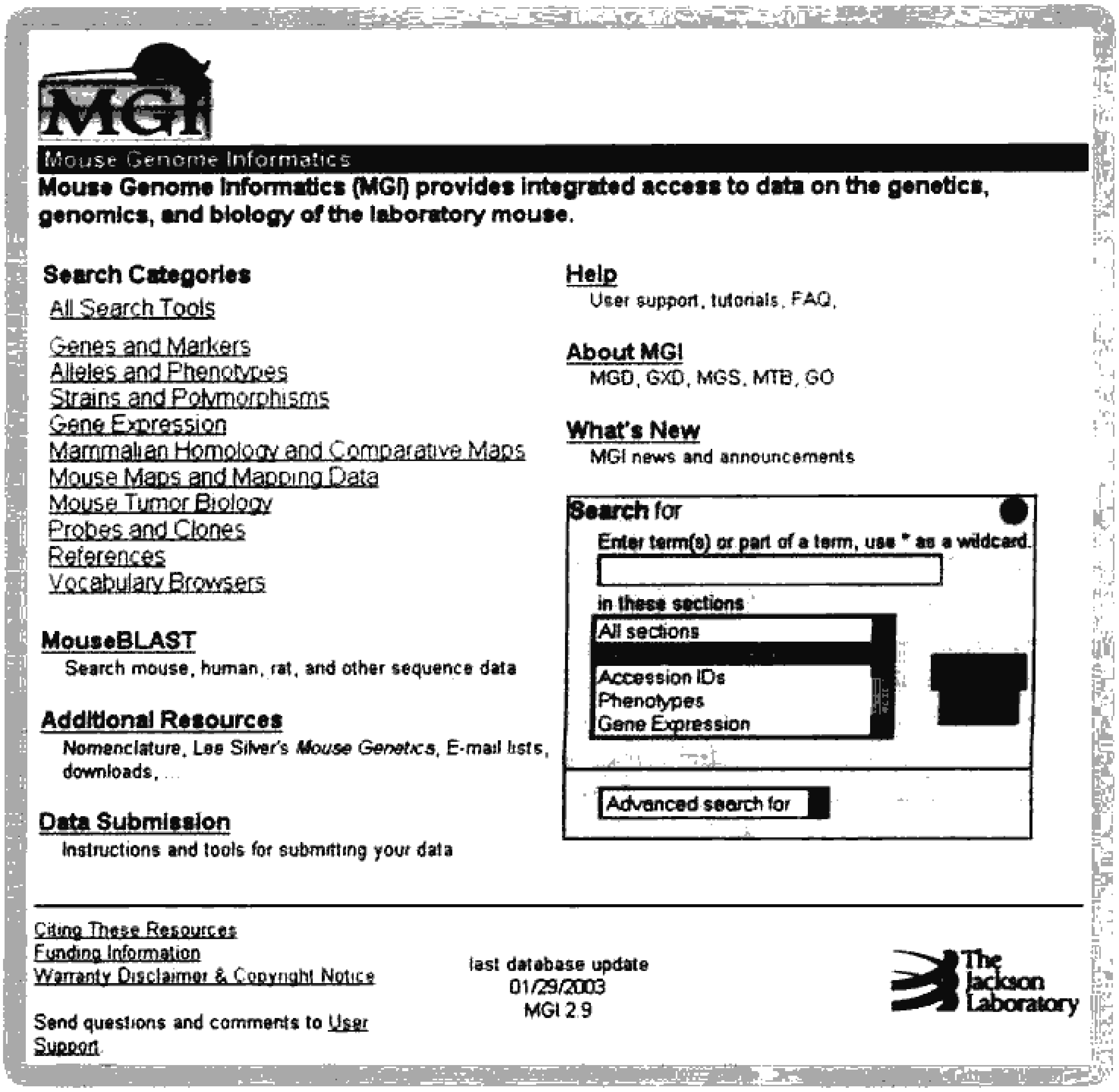


图 16-28 小鼠基因组信息数据库，由 Jackson 实验室开发，是关于小鼠资源的一个主要的在线数据库，它包括了许多的小鼠基因组数据

表 16-26 小鼠基因组资源

资源	注解	地址
Mouse genome project	Baylor College of Medicine	http://www.hgsc.bcm.tmc.edu/projects/mouse/
Mouse cytogenetic maps	Mammalian Genetics Unit, Harwell, United Kingdom	http://www.mgu.har.mrc.ac.uk/anomaly/anomaly-intro.html
TBASE	The Transgenic/Targeted Mutation Database	www.jax.org/tbase
Database of Gene Knockouts	At bioscience.org	http://www.bioscience.org/knockout/knochome.htm
Mouse Genetics	An online book by Lee Silver	http://www.informatics.jax.org/silver/
Mouse mapping	Genome Sciences Centre (Vancouver)	http://www.bcgsc.ca/lab/mapping/mouse
Trans-NIH Mouse Initiative	Comprehensive NIH mouse genomics site	http://www.nih.gov/science/models/mouse/
TIGR Mouse Gene Index	At TIGR	http://www.tigr.org/tdb/mgi/

表 16-27 大鼠基因组资源

资 源	注 解	地 址
Rat Genome Database	Key rat genomics site	http://rgd.mcw.edu
Ratmap	Rat Genome Database	http://ratmap.gen.gu.se/
NIH Rat Genomics and Genetics	Main NIH rat site	http://www.nih.gov/science/models/rat/
Rat Genome Resources	Central NCBI rat site	http://www.ncbi.nlm.nih.gov/genome/guide/rat/
TIGR Rat Gene Index	At TIGR	http://www.tigr.org/tdb/tgi/rgi/

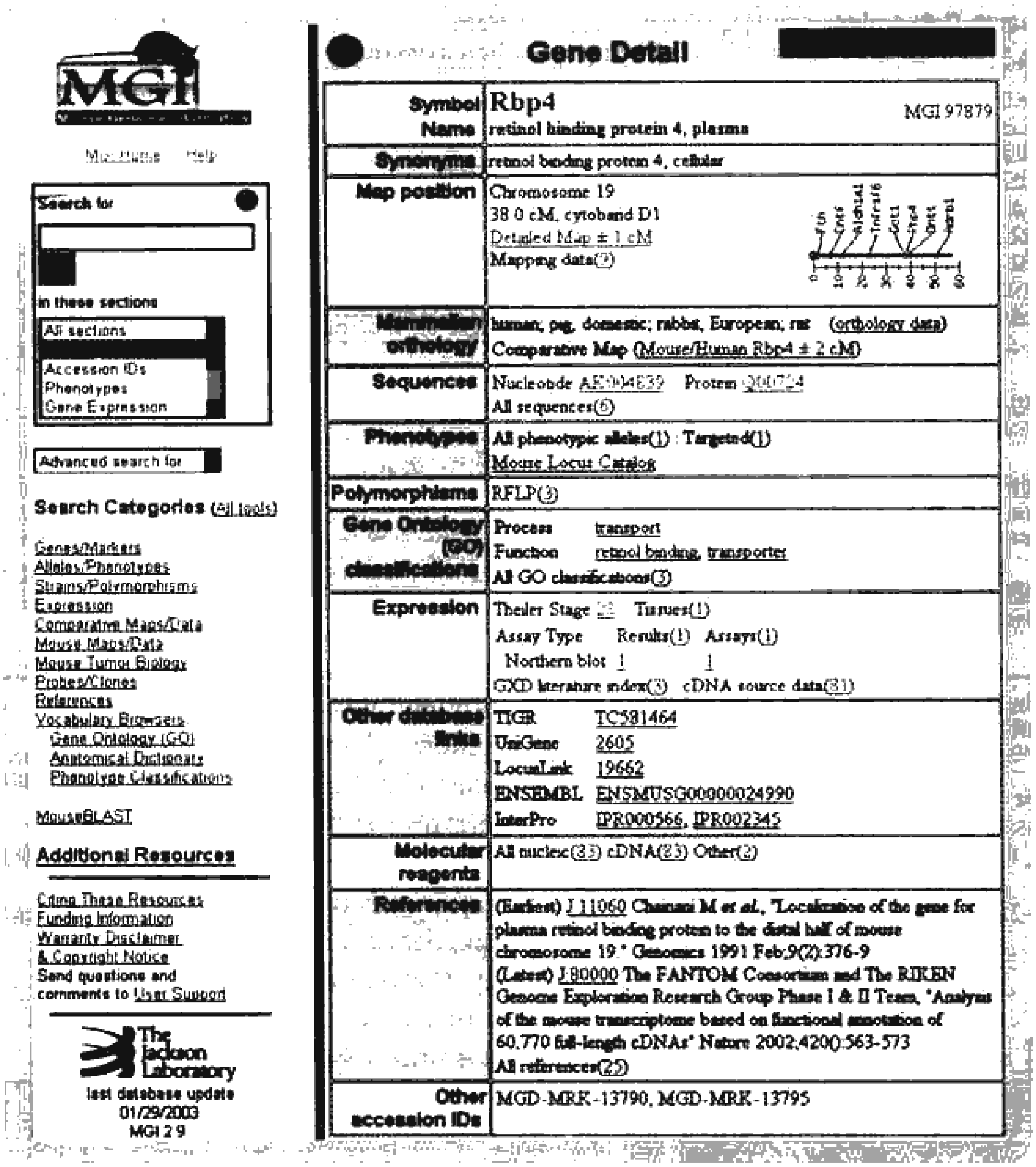


图 16-29 在小鼠基因组信息数据库中与 RBP 有关的结果包括了图谱链接，可能的直系同源蛋白、DNA 和蛋白质序列、多态性、GO 分类（见第 8 章）、表达谱、参考文献和其他数据

16.3.14 灵长类动物基因组

人类是从其他灵长类动物进化而来的吗？人类基因组中什么样的特性决定了我们独有的性状，如语言和高等智慧？多个灵长类动物的基因组比较将帮助理解我们特有性状的分子基础，或者说，将告诉我们在基因水平上我们与大型猿的亲缘关系。

我们可以从用溶解酶素构建一个种系发生树开始我们对灵长类动物的讨论（图 16-30）。黑猩猩和倭黑猩猩是与人类亲缘关系最近的两个物种。根据 36 个核内基因的序列分析

(Stauffer，2001)，这三个物种在大约（5.4±1.1）百万年前由一个共同祖先分化而来。其次同我们亲缘最近的是大猩猩，大约在（6.4±1.5）百万年前分化。分支树旁边是猩猩[(11.3±1.3）百万年前]和长臂猿[(14.9±2.0）百万年前]。类人猿（包括短尾猿和狒狒）大约在 0.23 亿年前从旧大陆猴中分化，其分化时间同发现的最早的类人猿化石形成时间相近。而新大陆猴（如绢毛猴）与人类的亲缘关系则更远。

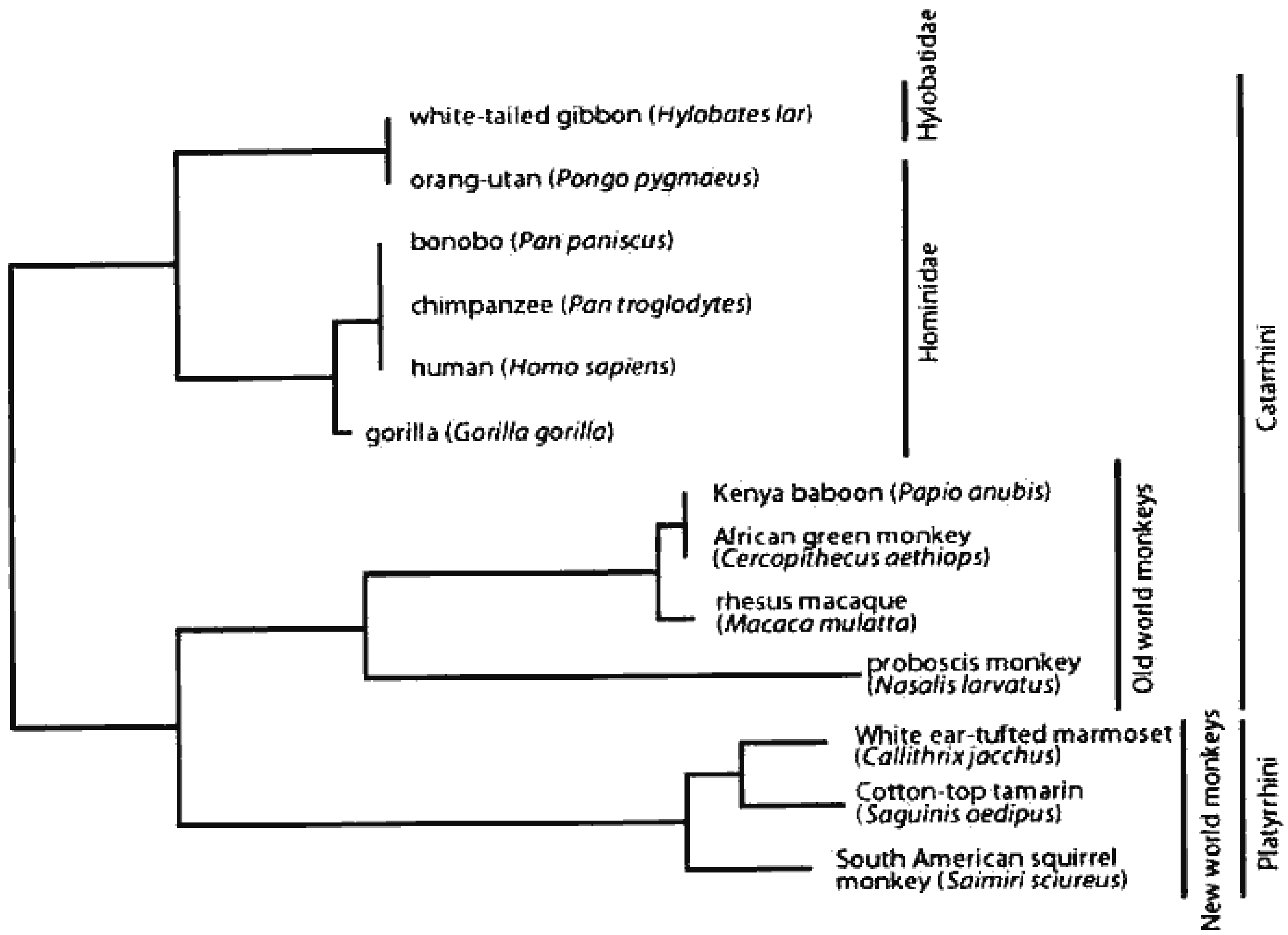


图 16-30　基于溶解酶蛋白质序列构建的灵长类动物种系 NJ 树。序列使用 ClustalW 并用 NJ 树表示。编号如下：长臂猿（P79180），猩猩（P79180），倭黑猩猩（AAB41214），黑猩猩（AAB41209），人类（P00695），大猩猩（P79179），肯尼亚倭黑猩猩（P00696），非洲绿猴（P00696），猕猴（P30201），长鼻猴（P79811），美洲产小型长尾猴（P79158），绢毛猴（P79268），南美洲松鼠猴（P79294）

成百上千个除人类外的灵长类动物的蛋白质和核酸序列已经被存储在 GenBank 中。对黑猩猩的大规模基因组测序计划正在进行之中。对黑猩猩的基因组分析将对人类的进化、遗传和疾病提供十分有用的信息。人类和黑猩猩 DNA 序列的 98％是一致的。

另两种即将被全基因组测序的动物是猕猴和狒狒。因为它们是亲缘关系较远的种属，它们比黑猩猩更频繁地被作为模式动物研究，关于它们的生理数据也更多。灵长类动物的资源参见表 16-28。

表 16-28　非人类灵长类基因组资源

资　源	注　解	地　址
Chimpanzee Genome Project	*Pan troglodytes* at Baylor College of Medicine	http://hgsc.bcm.tmc.edu/projects/chimpanzee/
Project Silver	National Institute of Genetics (Japan)	http://sayer.lab.nig.ac.jp/~silver/
Primate Cytogenetics Network	二倍体数目，核型和图谱	http://staff.washington.edu/timk/cyto/

非洲巨猿包括了黑猩猩和大猩猩。关于早期灵长类动物的进化（包括与树鼩、蝙蝠和飞狐猴的分化）请参见 Sargis（2002）。

16.4 展望

生物学的一个广泛目标是理解每种生命的本质：什么是发育、代谢、内稳态、繁殖和行为的机制？测定基因组序列并不能直接回答这些问题。相反，我们必须首先尝试对基因组进行注释，以便估计其包含的内容，然后我们试图在各种各样的生理过程中理解这些部分的功能。

在今后很短一段时间内，我们可以预料到所有主要真核生物种类的代表性生物的基因组将被测序。对于那些致病性的物种，我们希望基因组序列可以帮助我们理解它们毒性的细胞机制，它们逃避宿主免疫系统的机制，以及它们对于药物治疗的药理反应。对于进化研究来说，我们将更深入理解突变和选择等促使基因组进化的动力。

当全基因组序列被测定的时候，我们了解了非编码和编码 DNA 的特性。真核生物基因组的大部分都被重复序列所占据，包括那些转座子元件。编码蛋白质的基因数目从真菌的大约6000个到植物和哺乳动物的上万个。在每个物种中许多基因都是旁系同源的，因此对于许多真核生物来说，蛋白质组的核心部分大约是10000个基因数量级。在进化中，新的基因通过基因家族的扩增或对编码蛋白质结构域的 DNA 序列重组产生。

16.5 常见问题

目前基因组研究的一个迫切需求是继续开发能在基因组中找寻编码蛋白质基因、非编码RNA、重复序列、复制序列、保守共线性区域等的算法和软件。然后我们将在不同的发育阶段、不同的器官组织和不同的生理状态下确定基因的功能。通过这些方法，我们可以获得关于功能、进化和真核生物对环境的适应等方面的假说并加以检验。这样我们才能从基因组数据中获得其真正的含义。

我们目前处在基因组学的早期阶段，许多新的课题不断涌现。

① 基因组序列草图非常有用，但是当序列完成后基因注释经常得到改进。

② 在基因组 DNA 中预测蛋白质编码基因异常困难，特别是没有对应的关于基因表达的数据，如 EST 信息可用时，基因预测尤其困难。

③ 我们对非编码的 RNA 分子所知甚少。

④ 真核生物基因组中的大部分都是重复序列。

⑤ 比较基因组学在探索每个真核生物基因组的特性时非常有用。

许多描述真核生物或原核生物基因组的文献把从远古共同祖先的同一个基因分化出的后代基因当作直系同源基因。通常用 BLAST（E 值的阈值为 10^{-4}）对预测出来的蛋白质与其他生物的已知蛋白质组进行比对搜索。然而，两个直系同源蛋白在不同的物种中可能有不同的功能。

16.6 网络资源

我们在文中给出了许多真核生物基因组的网站资源。一个特别突出的例子是 Ensembl

(http://www.ensembl.org/)，目前它包括了小鼠、大鼠、斑马鱼、河豚、蚊子和其他生物的基因组。

16.7 问题讨论

1. 如果没有任何种类的重复 DNA，那么各种真核生物的基因组（人类、小鼠、植物和寄生虫的基因组）在大小、基因含量、基因顺序、核酸组成以及其他方面将有什么不同？

2. 如果有人给你来自于真核生物的 1Mb 基因组序列，你怎样鉴定它是属于哪个物种的（假定不能使用 BLAST 或者直接断定其物种）？哪些特性可以将原生生物、昆虫和鱼类的基因组 DNA 区分开来？

16.8 问题

1. (A) 从 Entrez 上获取一个拟南芥的 BAC 序列（如 BAC T18A20，GenBank 编号 AC009324）

(a) 注意其大小（kb）。它是一个大的 BAC 还是小的 BAC?

(b) 注意其中编码的蛋白质数目。细菌基因组中每千碱基对中有一个基因。在这个真核 DNA 中每千碱基对中有多少个基因？

(B) 在 NCBI 上找到 ORF Finder:

(a) 注意主页的侧栏，选择 "Tools for data mining"，然后你将看到 ORF Finder。

(b) 或者注意主页顶端的左侧栏，选择 "Site map"，然后就能看见一个到 ORF Finder 的链接。

(c) 把你的 BAC 编号粘贴进去，点击 OrfFind。

(C) 在 NCBI 的 ORF Finder 中点击最大的 ORF

(a) 它有多少个氨基酸？

(b) 分子量多少？

(c) 这个蛋白质是小蛋白质、中等蛋白质和大蛋白质？

(d) 这个基因是从 BAC 的哪条链转录的？另外的 ORF 是和它处于同一条链还是相反的链上？

(D) 用 NCBI 的 ORF Finder、BLAST 的默认参数搜索该 ORF

(a) 注意，结果页面不会自动重载，因此需要手动刷新页面。

(b) BLAST 结果中有许多拟南芥蛋白质。但是，如果使用这个 ORF 作为搜索序列，使用标准的 blastp 进行搜索，你将得到来自许多物种的匹配蛋白，你也会得到来自保守结构域数据库的一条匹配结果。因此，ORFFinder 中的 BLAST 工具不如一般的 BLAST 搜索全面。

2. 人类着丝粒通常含有长度为 171bp 的被称为 α 卫星序列的重复序列（编号 X07685），一共长达几千个碱基对。首先针对非冗余数据库进行 blastn 搜索。你观察到了什么类型的数据库结果？第二步，把你的 BLAST 搜索对象限制在非人类的数据（在 Entrez 的限制性选择框内敲入 "satellite NOT human [organism]. "）。搜索结果中有灵长类、啮齿类或者植物

的序列吗？为什么着丝粒序列会有这样的种系分布？你会认为每个物种都有它特有的着丝粒序列吗？

3. 一个广泛存在的微环结合蛋白（GenBank 编号 A54598）从锥体虫体内被提纯出来，它能感染昆虫。blastp 搜索显示它在植物、真菌和后生动物（比如线虫）中都有同源蛋白。这个蛋白在不同的物种中是如何命名的？它含有的结构域在保守结构域数据库中的名称是什么？

4. Leishmania major 中含有重复 DNA 序列（如编号 AF421497）。你如何找出它存在的范围？它定位在哪里（如某个特定的染色体或者某个染色体区域）？

16.9 自测题

1. *C* 值矛盾是

(a) 某些基因组中核苷酸 *C* 值较小；

(b) 真核生物基因组大小同其编码蛋白质的基因的个数没有相关性；

(c) 真核生物基因组大小同其物种的复杂度相关性很小；

(d) 真核生物基因组的大小同进化上的年龄相关性很小

2. 成百上千个以 4～8bp 为一个单位的重复序列经常出现在

(a) 散布性重复序列；

(b) 假基因；

(c) 端粒；

(d) 片段性复制区域

3. 当你正在测定一个新黏菌物种的基因组序列时，如果你使用 RepeatMasker 来探测其中的重复序列的组成，将会发生什么？(你使用的参数和用于人类基因组的参数相同)

(a) RepeatMasker 将成功鉴定出所有的重复 DNA 序列。两个物种间重复序列的相似度足够高，该软件可以用于你的黏菌 DNA。

(b) RepeatMasker 将成功鉴定出大部分的重复 DNA 序列，但是有一些重复序列是物种特异性的，因此有许多假阳性和假阴性结果。

(c) RepeatMasker 将不能成功鉴定出大部分的重复 DNA 序列，你需要在你的黏菌 DNA 序列上重新训练 RepeatMasker 算法。

(d) 不可能预测，因为不同物种的重复序列可能相似也可能不同

4. 在真核基因组 DNA 序列上使用从头预测的方法预测编码蛋白质的基因是很困难的，主要原因是什么？

(a) 外显子/内含子边界很难预测；

(b) 内含子长度可能有几个碱基对；

(c) 编码区域的 GC 含量并不总是与非编码区域的 GC 含量相同；

(d) 以上 3 点都是

5. 两种亲缘关系较远的真核生物基因组可能合并形成一个新的物种

(a) 正确；

(b) 错误

6. 贾第鞭毛虫基因组是非常特别的，因为

(a) 几乎不包含转座子元件或者内含子；

(b) 是环状的；

(c) 包含了很少的非重复 DNA；

(d) AT 含量接近 80%

7. 植物是地球上最成功的生物之一，包括了上万种物种。水稻和拟南芥基因组是最早被测序的两个植物基因组。将它们同后生动物的基因组进行广泛比较，则植物基因组

(a) GC 含量、基因组大小、基因数目

范围差不多；

(b) 不同，因为植物基因组的重复序列较少；

(c) 很可能不同，因为植物基因组维持了较快的进化速率；

(d) 根本不同，因为植物基因组使用了非典型的核苷酸酶

8. 线虫和酵母的基因组有何不同？

(a) 线虫基因组中编码多结构域蛋白的基因明显较多；

(b) 线虫的蛋白质核心部分比酵母的核心蛋白质多 10 倍；

(c) 线虫的 GC 含量明显较高；

(d) 线虫基因组中编码了没有已知同源物的蛋白质的基因比例较高

9. 小鼠和人类的基因组有什么不同？

(a) 小鼠基因组的 GC 含量更低；

(b) 小鼠基因组有更多的编码蛋白质的基因；

(c) 小鼠基因组中的某些特定基因家族（如嗅觉受体蛋白家族）有扩增现象；

(d) 小鼠基因组的每个染色体端粒的重复序列较少

16.10 推荐读物

Britten 和 Kohen (1968) 的关于重复性 DNA 序列的经典研究值得推荐。

对于真核生物基因预测，Makarov (2002) 的关于软件程序的综述文章值得推荐。

David Searls (2000) 在一章中对许多生物信息学工具的应用进行了综述（如 BLAST 和 ESTs）。

参考文献

Adam, R. D. Biology of *Giardia lamblia*. *Clin. Microbiol. Rev.* **14**, 447-475 (2001).

Adams, M. D., et al. The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185-2195 (2000).

Akman, L., et al. Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nat. Genet.* **32**, 402-407 (2002).

Allen, K. D. Assaying gene content in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **99**, 9568-9572 (2002).

Ambros, V. microRNAs: Tiny regulators with great potential. *Cell* **107**, 823-826 (2001).

Amor, D. J., and Choo, K. H. Neocentromeres: Role in human disease, evolution, and centromere study. *Am. J. Hum. Genet.* **71**, 695-714 (2002).

The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815 (2000).

Antonarakis, S. E., et al. Factor VIII gene inversions in severe hemophilia A: Results of an international consortium study. *Blood* **86**, 2206-2212 (1995).

Aparicio, S., et al. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301-1310 (2002).

Arkhipova, I., and Meselson, M. Transposable elements in sexual and ancient asexual taxa. *Proc. Natl. Acad. Sci. USA* **97**, 14473-14477 (2000).

Arkhipova, I. R., and Morrison, H. G. Three retrotransposon families in the genome of *Giardia lamblia*: Two telomeric, one dead. *Proc. Natl. Acad. Sci. USA* **98**, 14497-14502 (2001).

Avramova, Z. V. Heterochromatin in animals and plants. Similarities and differences. *Plant Physiol.* **129**, 40-49 (2002).

Azzalin, C. M., Nergadze, S. G., and Giulotto, E. Human intrachromosomal telomeric-like repeats: Sequence organization and mechanisms of origin. *Chromosoma* **110**, 75-82 (2001).

Bahl, A., et al. PlasmoDB: The *Plasmodium* genome resource. A database integrating experimental and computational data. *Nucleic Acids Res.* **31**, 212-215 (2003).

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., and Eichler, E. E. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005-1017 (2001).

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., and Doolittle, W. F. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290**, 972-977 (2000).

Bennetzen, J. L., and Freeling, M. The unified grass genome: Synergy in synteny. *Genome Res.* **7**, 301-306 (1997).

Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573-580 (1999).

Ben-Yehuda, S., Rudner, D. Z., and Losick, R. RacA, a bacterial protein that anchors chromosomes to the cell poles. *Science* **19**, 19 (2002).

Betrán, E., and Long, M. Expansion of genome coding regions by acquisition of new genes. *Genetica* **115**, 65-80 (2002).

Blackburn, E. H., et al. Recognition and elongation of telomeres by telomerase. *Genome* **31**, 553-560 (1989).

Blake, J. A., Richardson, J. E., Bult, C. J., Kadin, J. A., and Eppig, J. T. MGD: The Mouse Genome Database. *Nucleic Acids Res.* **31,** 193–195 (2003).

Blattner, F. R., et al. The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–1474 (1997).

Blaxter, M. *Caenorhabditis elegans* is a nematode. *Science* **282,** 2041–2046 (1998).

Blaxter, M. Molecular systematics: Counting angels with DNA. *Nature* **421,** 122–124 (2003).

Bray, N., Dubchak, I., and Pachter, L. AVID: A Global Alignment Program. *Genome Res.* **13,** 97–102 (2003).

Britten, R. J., and Kohne, D. E. Repeated sequences in DNA. *Science* **161,** 529–540 (1968).

Brooks, D. R., and Isaac, R. E. Functional genomics of parasitic worms: The dawn of a new era. *Parasitol. Int.* **51,** 319–325 (2002).

Budiansky, S. Creatures of our own making. *Science* **298,** 80–86 (2002).

Buell, C. R. Current status of the sequence of the rice genome and prospects for finishing the first monocot genome. *Plant Physiol.* **130,** 1585–1586 (2002).

Burge, C., and Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).

Burge, C. B., and Karlin, S. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8,** 346–354 (1998).

Cameron, R. A., et al. A sea urchin genome project: Sequence scan, virtual map, and additional resources. *Proc. Natl. Acad. Sci. USA* **97,** 9514–9518 (2000).

Carlton, J. M., et al. Profiling the malaria genome: A gene survey of three species of malaria parasite with comparison to other apicomplexan species. *Mol. Biochem. Parasitol.* **118,** 201–210 (2001).

Carlton, J. M., et al. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii.* *Nature* **419,** 512–519 (2002).

Cavalier-Smith, T. A revised six-kingdom system of life. *Biol. Rev. Camb. Philos. Soc.* **73,** 203–266 (1998).

Cavalier-Smith, T. Origins of the machinery of recombination and sex. *Heredity* **88,** 125–141 (2002).

Celniker, S. E., et al. Finishing a whole-genome shotgun: Release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* **3,** RESEARCH0079-9 (2002).

Centers for Disease Control and Prevention. Provisional surveillance summary of the West Nile virus epidemic—United States, January-November 2002. *MMWR* **51,** 1129–1133 (2002).

Chan, P., Simon-Chazottes, D., Mattei, M. G., Guenet, J. L., and Salier, J. P. Comparative mapping of lipocalin genes in human and mouse: The four genes for complement C8 gamma chain, prostaglandin-D-synthase, oncogene-24p3, and progestagen-associated endometrial protein map to HSA9 and MMU2. *Genomics* **23,** 145–150 (1994).

Choo, K. H. Domain organization at the centromere and neocentromere. *Dev. Cell.* **1,** 165–177 (2001).

Claverie, J. M. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6,** 1735–1744 (1997).

Claverie, J. M. Gene number. What if there are only 30,000 human genes? *Science* **291,** 1255–1257 (2001).

Comai, L. Genetic and epigenetic interactions in allopolyploid plants. *Plant Mol. Biol.* **43,** 387–399 (2000).

Cowman, A. F., and Crabb, B. S. The *Plasmodium falciparum* genome—a blueprint for erythrocyte invasion. *Science* **298,** 126–128 (2002).

Cox, F. E. History of human parasitology. *Clin. Microbiol. Rev.* **15,** 595–612 (2002).

Crowe, M. L., Serizet, C., Thareau, V., Aubourg, S., Rouze, P., Hilson, P., Beynon, J., Weisbeek, P., van Hummelen, P., Reymond, P., Paz-Ares, J., Nietfeld, W., Trick, M. CATMA: a complete *Arabidopsis* GST database. *Nucleic Acids Res.* **31,** 156–158 (2003).

Cummings, C. J., and Zoghbi, H. Y. Trinucleotide repeats: Mechanisms and pathophysiology. *Annu. Rev. Genomics Hum. Genet.* **1,** 281–328 (2000).

Cuvier, G. *The Animal Kingdom, Arranged According to its Organization.* William S. Orr & Co., London, 1849.

Dehal, P., et al. Human chromosome 19 and related regions in mouse: Conservative and lineage-specific evolution. *Science* **293,** 104–111 (2001).

Dehal, P., et al. The draft genome of *Ciona intestinalis:* Insights into chordate and vertebrate origins. *Science* **298,** 2157–2167 (2002).

Dermitzakis, E. T., et al. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420,** 578–582 (2002).

Dessen, P., et al. Paramecium genome survey: A pilot project. *Trends Genet.* **17,** 306–308 (2001).

Dewald, G., et al. The human complement C8G gene, a member of the lipocalin gene family: Polymorphisms and mapping to chromosome 9q34.3. *Ann. Hum. Genet.* **60,** 281–291 (1996).

Donelson, J. E. Genome research and evolution in trypanosomes. *Curr. Opin. Genet. Dev.* **6,** 699–703 (1996).

Dong, Q., Roy, L., Freeling, M., Walbot, V., and Brendel, V. ZmDB, an integrated database for maize genome research. *Nucleic Acids Res.* **31,** 244–247 (2003).

Douglas, S. E., and Penny, S. L. The plastid genome of the cryptophyte alga, *Guillardia theta:* Complete sequence and conserved synteny groups confirm its common ancestry with red algae. *J. Mol. Evol.* **48,** 236–244 (1999).

Douglas, S., et al. The highly reduced genome of an enslaved algal nucleus. *Nature* **410,** 1091–1096 (2001).

Echols, N., et al. Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. *Nucleic Acids Res.* **30,** 2515–2523 (2002).

Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2,** 919–929 (2001).

Eddy, S. R. Computational genomics of noncoding RNA genes. *Cell* **109,** 137–140 (2002).

El-Sayed, N. M., Hegde, P., Quackenbush, J., Melville, S. E., and Donelson, J. E. The African trypanosome genome. *Int. J. Parasitol.* **30,** 329–345 (2000).

Embley, T. M., and Hirt, R. P. Early branching eukaryotes? *Curr. Opin. Genet. Dev.* **8,** 624–629 (1998).

Fan, Y., Linardopoulou, E., Friedman, C., Williams, E., and Trask, B. J. Genomic structure and evolution of the ancestral chromosome fusion site in 2q13–2q14.1 and paralogous regions on other human chromosomes. *Genome Res.* **12,** 1651–1662 (2002).

Feng, Q., et al. Sequence and analysis of rice chromosome 4.

Nature **420**, 316–320 (2002).

Fields, S., Kohara, Y., and Lockhart, D. J. Functional genomics. *Proc. Natl. Acad. Sci. USA* **96**, 8825–8256 (1999).

Florens, L., et al. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520–526 (2002).

The FlyBase Consortium. The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.* **31**, 172–175 (2003).

Fraser, C. M., Eisen, J. A., and Salzberg, S. L. Microbial genome sequencing. *Nature* **406**, 799–803 (2000).

Frazer, K. A., Elnitski, L., Church, D. M., Dubchak, I., and Hardison, R. C. Cross-species sequence comparisons: A review of methods and available resources. *Genome Res.* **13**, 1–12 (2003).

Gardner, M. J., et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).

Gelbart, W. M. Databases in genomic research. *Science* **282**, 659–661 (1998).

The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**, 2012–2018 (1998).

Gendles, A. J., and Karlin, S. Genome-scale compositional comparisons in eukaryotes. *Genome Res.* **11**, 540–546 (2001).

Gilson, P. R., and McFadden, G. I. A grin without a cat. *Nature* **410**, 1040–1041 (2001).

Gilson, P. R., and McFadden, G. I. Jam packed genomes—a preliminary, comparative analysis of nucleomorphs. *Genetica* **115**, 13–28 (2002).

Glockner, G., et al. Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. *Nature* **418**, 79–85 (2002).

Goff, S. A., et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92–100 (2002).

Graur, D., and Li, W.-H. *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA, 2000.

Gregory, S. G., et al. A physical map of the mouse genome. *Nature* **418**, 743–750 (2002).

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R. Rfam: An RNA family database. *Nucleic Acids Res.* **31**, 439–441 (2003).

Gutierrez, G., Ganfornina, M. D., and Sanchez, D. Evolution of the lipocalin family as inferred from a protein sequence phylogeny. *Biochim. Biophys. Acta* **1482**, 35–45 (2000).

Hall, A. E., Fiebig, A., and Preuss, D. Beyond the *Arabidopsis* genome: Opportunities for comparative genomics. *Plant Physiol.* **129**, 1439–1447 (2002).

Hancock, J. M. Genome size and the accumulation of simple sequence repeats: Implications of new data from genome sequencing projects. *Genetica* **115**, 93–103 (2002).

Hardison, R. C., Oeltjen, J., and Miller, W. Long human-mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7**, 959–966 (1997).

Harris, T. W., et al. WormBase: A cross-species database for comparative genomics. *Nucleic Acids Res.* **31**, 133–137 (2003).

Harrison, P. M., and Gerstein, M. Studying genomes through the aeons: Protein families, pseudogenes and proteome evolution. *J. Mol. Biol.* **318**, 1155–1174 (2002).

Harrison, P. M., Kumar, A., Lang, N., Snyder, M., and Gerstein, M. A question of size: The eukaryotic proteome and the problems in defining it. *Nucleic Acids Res.* **30**, 1083–1090 (2002).

Hartl, D. L. Molecular melodies in high and low C. *Nat. Rev. Genet.* **1**, 145–149 (2000).

Hebert, P. D. N., Cywinska, A., Ball, S. L., and deWaard, J. R. Biological identification through DNA barcodes. *Proc. R. Soc. Lond.* **270**, 313–321 (2003).

Hedges, S. B., and Kumar, S. Genomics. Vertebrate genomes compared. *Science* **297**, 1283–1285 (2002).

Henrich, T., et al. MEPD: A Medaka gene expression pattern database. *Nucleic Acids Res.* **31**, 72–74 (2003).

Hill, A. A., Hunter, C. P., Tsung, B. T., Tucker-Kellogg, G., and Brown, E. L. Genomic analysis of gene expression in *C. elegans*. *Science* **290**, 809–812 (2000).

Hoffman, S. L., Subramanian, G. M., Collins, F. H., and Venter, J. C. *Plasmodium*, human and *Anopheles* genomics and malaria. *Nature* **415**, 702–709 (2002).

Holland, P. W. Ciona. *Curr. Biol.* **12**, R609 (2002).

Holt, R. A., et al. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129–149 (2002).

Hou, G., Le Blancq, S. M., Yaping, E., Zhu, H., and Lee, M. G. Structure of a frequently rearranged rRNA-encoding chromosome in *Giardia lamblia*. *Nucleic Acids Res.* **23**, 3310–3317 (1995).

IJdo, J. W., Baldini, A., Ward, D. C., Reeders, S. T., and Wells, R. A. Origin of human chromosome 2: An ancestral telomere-telomere fusion. *Proc. Natl. Acad. Sci. USA* **88**, 9051–9055 (1991a)

IJdo, J. W., Wells, R. A., Baldini, A., and Reeders, S. T. Improved telomere detection using a telomere repeat probe (TTAGGG)n generated by PCR. *Nucleic Acids Res.* **19**, 4780 (1991b).

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

Jarstfer, M. B., and Cech, T. R. Effects of nucleotide analogues on *Euplotes aediculatus* telomerase processivity: Evidence for product-assisted translocation. *Biochemistry* **41**, 151–161 (2002).

Jentsch, S., Tobler, H., and Muller, F. New telomere formation during the process of chromatin diminution in *Ascaris suum*. *Int. J. Dev. Biol.* **46**, 143–148 (2002).

Johnson, P. J. Spliceosomal introns in a deep-branching eukaryote: The splice of life. *Proc. Natl. Acad. Sci. USA* **99**, 3359–3361 (2002).

Jomaa, H., et al. Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science* **285**, 1573–1576 (1999).

Jones, K. L. *Smith's Recognizable Patterns of Human Malformation*, W. B. Saunders, New York, 1997.

Jurka, J. Repeats in genomic DNA: Mining and meaning. *Curr. Opin. Struct. Biol.* **8**, 333–337 (1998).

Jurka, J. Repbase update: A database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).

Kamath, R. S., et al. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237 (2003).

Kapranov, P., et al. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).

Karlowski, W. M., Schoof, H., Janakiraman, V., Stuempflen, V., and Mayer, K. F. MOsDB: An integrated information resource for rice genomics. *Nucleic Acids Res.* **31**, 190–192

(2003).

Kidwell, M. G. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**, 49–63 (2002).

Kim, S. K. Functional genomics: The worm scores a knockout. *Curr. Biol.* **11**, R85–R87 (2001).

Kissinger, J. C., Gajria, B., Li, L., Paulsen, I. T., and Roos, D. S. ToxoDB: Accessing the *Toxoplasma gondii* genome. *Nucleic Acids Res.* **31**, 234–236 (2003).

Knight, J. All genomes great and small. *Nature* **417**, 374–376 (2002).

Knight, R. D., Freeland, S. J., and Landweber, L. F. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* **2** (2001).

Kotzot, D. Complex and segmental uniparental disomy (UPD): Review and lessons from rare chromosomal complements. *J. Med. Genet.* **38**, 497–507 (2001).

Ku, H. M., Vision, T., Liu, J., and Tanksley, S. D. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci. USA* **97**, 9121–9126 (2000).

Kumar, S., and Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).

Lacazette, E., Gachon, A. M., and Pitiot, G. A novel human odorant-binding protein gene family resulting from genomic duplicons at 9q34: Differential expression in the oral and genital spheres. *Hum. Mol. Genet.* **9**, 289–301 (2000).

Lamblin, A. F., et al. MtDB: A database for personalized data mining of the model legume *Medicago truncatula* transcriptome. *Nucleic Acids Res.* **31**, 196–201 (2003).

Lasonder, E., et al. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419**, 537–542 (2002).

Lee, H. S., and Chen, Z. J. Protein-coding genes are epigenetically regulated in *Arabidopsis* polyploids. *Proc. Natl. Acad. Sci. USA* **98**, 6753–6758 (2001).

Lewis, M. S., and Pikaard, C. S. Restricted chromosomal silencing in nucleolar dominance. *Proc. Natl. Acad. Sci. USA* **98**, 14536–14540 (2001).

Lloyd, D., and Harris, J. C. *Giardia:* Highly evolved parasite or early branching eukaryote? *Trends Microbiol.* **10**, 122–127 (2002).

Long, C. A., and Hoffman, S. L. Malaria—from infants to genomics to vaccines. *Science* **297**, 345–347 (2002).

Lorite, P., Carrillo, J. A., and Palomeque, T. Conservation of (TTAGG)(n) telomeric sequences among ants (Hymenoptera, Formicidae). *J. Hered.* **93**, 282–285 (2002).

Lowe, T. M., and Eddy, S. R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).

Makalowski, W. Genomic scrap yard: How genomes utilize all that junk. *Gene* **259**, 61–67 (2000).

Makarov, V. Computer programs for eukaryotic gene prediction. *Brief. Bioinform.* **3**, 195–199 (2002).

Margulis, L., and Schwartz, K. V. *Five Kingdoms. An Illustrated Guide to the Phyla of Life on Earth.* W. H. Freeman, New York, 1998.

Martin, C. L., et al. The evolutionary origin of human subtelomeric homologies—or where the ends begin. *Am. J. Hum. Genet.* **70**, 972–984 (2002).

Matthews, D. E., Carollo, V. L., Lazo, G. R., and Anderson, O. D. GrainGenes, the genome database for small-grain crops. *Nucleic Acids Res.* **31**, 183–186 (2003).

Mayor, C., et al. VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**, 1046–1047 (2000).

McArthur, A. G., et al. The *Giardia* genome project database. *FEMS Microbiol. Lett.* **189**, 271–273 (2000).

McConkey, E. H., and Varki, A. A primate genome project deserves high priority. *Science* **289**, 1295–1296 (2000).

McCormick-Graham, M., and Romero, D. P. A single telomerase RNA is sufficient for the synthesis of variable telomeric DNA repeats in ciliates of the genus *Paramecium*. *Mol. Cell. Biol.* **16**, 1871–1879 (1996).

McKnight, T. D., Fitzgerald, M. S., and Shippen, D. E. Plant telomeres and telomerases. A review. *Biochemistry (Mosc.)* **62**, 1224–1231 (1997).

Melek, M., Davis, B. T., and Shippen, D. E. Oligonucleotides complementary to the *Oxytricha nova* telomerase RNA delineate the template domain and uncover a novel mode of primer utilization. *Mol. Cell. Biol.* **14**, 7827–7838 (1994).

Meyerowitz, E. M. Plants compared to animals: The broadest comparative study of development. *Science* **295**, 1482–1485 (2002).

Misra, S., et al. Annotation of the *Drosophila melanogaster* euchromatic genome: A systematic review. *Genome Biol.* **3**, RESEARCH0083-3 (2002).

Moller-Jensen, J., Jensen, R. B., Lowe, J., and Gerdes, K. Prokaryotic DNA segregation by an actin-like filament. *EMBO J.* **21**, 3119–3127 (2002).

Morgenstern, B., et al. Exon discovery by genomic sequence alignment. *Bioinformatics* **18**, 777–787 (2002).

Morris, J. C., et al. Replication of kinetoplast DNA: An update for the new millennium. *Int. J. Parasitol.* **31**, 453–458 (2001).

Müller, F., and Tobler, H. Chromatin diminution in the parasitic nematodes *Ascaris suum* and *Parascaris univalens*. *Int. J. Parasitol.* **30**, 391–399 (2000).

Mural, R. J. Current status of computational gene finding: A perspective. *Methods Enzymol.* **303**, 77–83 (1999).

Mural, R. J., et al. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296**, 1661–1671 (2002).

Murthy, V. L., and Rose, G. D. RNABase: An annotated database of RNA structures. *Nucleic Acids Res.* **31**, 502–504 (2003).

Myler, P. J., et al. *Leishmania major* Friedlin chromosome 1 has an unusual distribution of protein-coding genes. *Proc. Natl. Acad. Sci. USA* **96**, 2902–2906 (1999).

Myler, P. J., et al. Genomic organization and gene function in *Leishmania*. *Biochem. Soc. Trans.* **28**, 527–531 (2000).

Nadeau, J. H., and Taylor, B. A. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. USA* **81**, 814–818 (1984).

Nanda, I., et al. Distribution of telomeric (TTAGGG)(n) sequences in avian chromosomes. *Chromosoma* **111**, 215–227 (2002).

Nixon, J. E., et al. A spliceosomal intron in *Giardia lamblia*. *Proc. Natl. Acad. Sci. USA* **99**, 3701–3705 (2002).

Novichkov, P. S., Gelfand, M. S., and Mironov, A. A. Gene

recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics* **17**, 1011–1018 (2001).

Ostertag, E. M., and Kazazian, H. H., Jr. Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* **11**, 2059–2065 (2001).

Pardue, M. L., DeBaryshe, P. G., and Lowenhaupt, K. Another protozoan contributes to understanding telomeres and transposable elements. *Proc. Natl. Acad. Sci. USA* **98**, 14195–14197 (2001).

Passarge, E., Horsthemke, B., and Farber, R. A. Incorrect use of the term synteny. *Nat. Genet.* **23**, 387 (1999).

Pelosi, P. Perireceptor events in olfaction. *J. Neurobiol.* **30**, 3–19 (1996).

Pevsner, J., Hou, V., Snowman, A. M., and Snyder, S. H. Odorant-binding protein. Characterization of ligand binding. *J. Biol. Chem.* **265**, 6118–6125 (1990).

Philippe, H., and Laurent, J. How good are deep phylogenetic trees? *Curr. Opin. Genet. Dev.* **8**, 616–623 (1998).

Redi, C. A., Garagna, S., Zacharias, H., Zuccotti, M., and Capanna, E. The other chromatin. *Chromosoma* **110**, 136–147 (2001).

Reese, M. G., Kulp, D., Tammana, H., and Haussler, D. Genie—gene finding in *Drosophila melanogaster*. *Genome Res.* **10**, 529–538 (2000).

Reich, D. E., Lander, E. S., Waterson, R., Pääbo, S., Ruvolo, M., and Varki, A. Sequencing the chimpanzee genome. On-line white paper (not dated), ▸www.hgsc.bcm.tmc.edu/projects/chimpanzee/ChimpGenome2.pdf.

Reyes, A., Pesole, G., and Saccone, C. Long-branch attraction phenomenon and the impact of among-site rate variation on rodent phylogeny. *Gene* **259**, 177–187 (2000).

Rhee, S. Y., et al. The Arabidopsis Information Resource (TAIR): A model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.* **31**, 224–228 (2003).

Rogic, S., Mackworth, A. K., and Ouellette, F. B. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* **11**, 817–832 (2001).

Rubin, G. M., and Lewis, E. B. A brief history of *Drosophila*'s contributions to genome research. *Science* **287**, 2216–2218 (2000).

Rubin, G. M., et al. Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).

Rudd, S., Mewes, H. W., and Mayer, K. F. Sputnik: A database platform for comparative plant genomics. *Nucleic Acids Res.* **31**, 128–132 (2003).

Ruvkun, G. Molecular biology. Glimpses of a tiny RNA world. *Science* **294**, 797–799 (2001).

Ryan, J. F., and Finnerty, J. R. CnidBase: The Cnidarian Evolutionary Genomics Database. *Nucleic Acids Res.* **31**, 159–163 (2003).

Salier, J. P. Chromosomal location, exon/intron organization and evolution of lipocalin genes. *Biochim. Biophys. Acta* **1482**, 25–34 (2000).

Samson, D., et al. GenoPlante-Info (GPI): A collection of databases and bioinformatics resources for plant genomics. *Nucleic Acids Res.* **31**, 179–182 (2003).

Sargis, E. J. Paleontology. Primate origins nailed. *Science* **298**, 1564–1565 (2002).

Sasaki, T., et al. The genome sequence and structure of rice chromosome 1. *Nature* **420**, 312–316 (2002).

Schlötterer, C., and Harr, B. *Drosophila virilis* has long and highly polymorphic microsatellites. *Mol. Biol. Evol.* **17**, 1641–1646 (2000).

Schwartz, S., et al. PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.* **10**, 577–586 (2000).

Searls, D. B. *Bioinformatics Tools for Whole Genomes.* Annual Reviews, Palo Alto, CA, 2000, pp. 251–279.

Seo, H. C., et al. Miniature genome in the marine chordate *Oikopleura dioica*. *Science* **294**, 2506 (2001).

Shapiro, T. A., and Englund, P. T. The structure and replication of kinetoplast DNA. *Annu. Rev. Microbiol.* **49**, 117–143 (1995).

Shippen-Lentz, D., and Blackburn, E. H. Telomere terminal transferase activity from *Euplotes crassus* adds large numbers of TTTTGGGG repeats onto telomeric primers. *Mol. Cell. Biol.* **9**, 2761–2764 (1989).

Simillion, C., Vandepoele, K., Van Montagu, M. C., Zabeau, M., and Van de Peer, Y. The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **99**, 13627–13632 (2002).

Simpson, A. G., MacQuarrie, E. K., and Roger, A. J. Eukaryotic evolution: Early origin of canonical introns. *Nature* **419**, 270 (2002).

Slijepcevic, P. Telomeres and mechanisms of Robertsonian fusion. *Chromosoma* **107**, 136–140 (1998a).

Slijepcevic, P. Telomere length regulation—a view from the individual chromosome perspective. *Exp. Cell Res.* **244**, 268–274 (1998b).

Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9**, 657–663 (1999).

Snel, B., Bork, P., and Huynen, M. A. Genome phylogeny based on gene content. *Nat. Genet.* **21**, 108–110 (1999).

Sonnhammer, E. L., and Durbin, R. Analysis of protein domain families in *Caenorhabditis elegans*. *Genomics* **46**, 200–216 (1997).

Sprague, J., et al. The Zebrafish Information Network (ZFIN): The zebrafish model organism database. *Nucleic Acids Res.* **31**, 241–243 (2003).

Stauffer, R. L., Walker, A., Ryder, O. A., Lyons-Weiler, M., and Hedges, S. B. Human and ape molecular clocks and constraints on paleontological hypotheses. *J. Hered.* **92**, 469–474 (2001).

Stechmann, A., and Cavalier-Smith, T. Rooting the eukaryote tree by using a derived gene fusion. *Science* **297**, 89–91 (2002).

Stein, L. Genome annotation: From sequence to biology. *Nat. Rev. Genet.* **2**, 493–503 (2001).

Szymanski, M., Erdmann, V. A., and Barciszewski, J. Noncoding regulatory RNAs database. *Nucleic Acids Res.* **31**, 429–431 (2003).

Tzafrir, I., et al. The *Arabidopsis* SeedGenes Project. *Nucleic Acids Res.* **31**, 90–93 (2003).

Upcroft, P., Chen, N., and Upcroft, J. A. Telomeric organization of a variable and inducible toxin gene family in the ancient eukaryote *Giardia duodenalis*. *Genome Res* **7**, 37–46 (1997).

Van de Peer, Y., Baldauf, S. L., Doolittle, W. F., and Meyer, A. An updated and comprehensive rRNA phylogeny of (crown) eukaryotes based on rate-calibrated evolutionary distances. *J. Mol. Evol.* **51**, 565–576 (2000).

Vellai, T., and Vida, G. The origin of eukaryotes: The difference

between prokaryotic and eukaryotic cells. *Proc. R. Soc. Lond. B Biol. Sci.* **266,** 1571–1577 (1999).

Walbot, V. *Arabidopsis thaliana* genome. A green chapter in the book of life. *Nature* **408,** 794–795 (2000).

Wang, D. Y., Kumar, S., and Hedges, S. B. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B Biol. Sci.* **266,** 163–171 (1999).

Waterston, R. H., et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).

Watt, W. B., and Dean, A. M. Molecular-functional studies of adaptive genetic variation in prokaryotes and eukaryotes. *Annu. Rev. Genet.* **34,** 593–622 (2000).

Williams, B. A., Hirt, R. P., Lucocq, J. M., and Embley, T. M. A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*. *Nature* **418,** 865–869 (2002).

Wirth, D. F. Biological revelations. *Nature* **419,** 495–496 (2002).

Wolfe, K. H., Morden, C. W., and Palmer, J. D. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc. Natl. Acad. Sci. USA* **89,** 10648–10652 (1992).

Wu, J., et al. A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell* **14,** 525–535 (2002).

Xuan, Z., McCombie, W. R., and Zhang, M. Q. GFScan: A gene family search tool at genomic DNA level. *Genome Res.* **12,** 1142–1149 (2002).

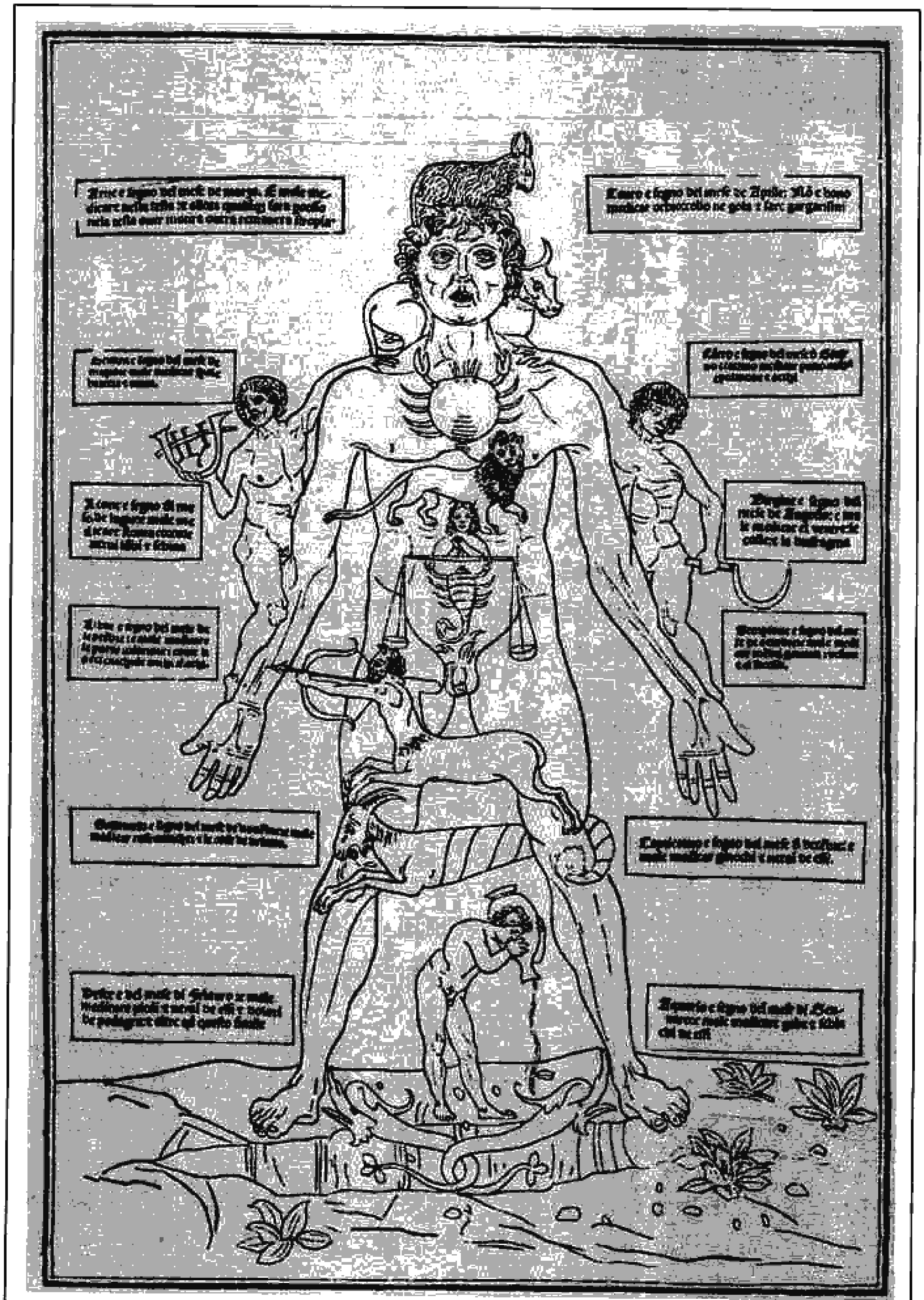
Xuan, Z., Wang, J., and Zhang, M. Q. Computational comparison of two mouse draft genomes and the human golden path. *Genome Biol.* **4** (2003).

Yu, J., et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296,** 79–92 (2002).

Yuan, Q., et al. The TIGR rice genome annotation resource: Annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Res.* **31,** 229–233 (2003).

Zdobnov, E. M., et al. Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298,** 149–159 (2002).

Zuegge, J., Ralph, S., Schmuker, M., McFadden, G. I., and Schneider, G. Deciphering apicoplast targeting signals—feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. *Gene* **280,** 19–26 (2001).



来自 Johannes de Ketham 所著“Fasciculus medicinae”（1493；参见 Singer，1925）的十二天宫人体图。图中将人体的各个部分用星座标记。在好几个世纪里，这种图被用来作为放血术的指导手册。将人体的器官与天上的星座对应起来，当月亮出现在病人器官的标记中时，医生避免在此时进行治疗。在图的左上方，第一个标注写道：“白羊座是三月的标记，这时不能进行头部的治疗手术，在那时头部受伤的人，要么死去，要么保持受伤的状态”；“双子座是五月的标志，这时不能进行肩部、手臂和手的放血手术。”

第 17 章

人类基因组

17.1 引言

人类基因组是智人［*Homo sapiens*］DNA 的完整集合。这些 DNA 编码的蛋白决定了我们作为生物实体的存在。通过基因组 DNA，编码蛋白基因得以表达，从而构成我们身体万亿个细胞的有序体系。而人与人之间的不同，从身体特征、个性人格到患病状况，也全都来源于基因组的差异。

人类基因组得以测序是科学的巨大成功。这个计划的完成恰好在 Crick 和 Watson 发表 DNA 双螺旋结构（1953）的 50 年之后。测序依赖于数百名科学家参与的国际协作［公众基金资助的国际人类基因组测序联盟（IHGSC），见下文］，同时，如果没有新兴的生物信息学和基因组学领域的基础性进步，测序计划也将难以进行。

本章分为三个部分。首先，我们总结了人类基因组计划的一些主要发现。然后介绍人类基因组研究的网络资源，涉及三个站点：美国国立生物技术信息中心（NCBI）、Ensembl 软件系统（the Ensembl project），以及加州大学圣塔克鲁兹分校（UCSC）的基因组研究中心。

在本章的第三部分将根据 2001 年 2 月 15 日国际人类基因组测序联盟发表的 62 页的人类基因组测序及分析草图文章的脉络，从生物信息学的角度来讲述人类基因组。我们也会提到塞莱拉公司（Celera Genomics）版本的人类基因组草图（Venter 等，2001）和他们的发现。

在网络上探索人类基因组可以从几个主要的站点开始。美国国家人类基因组研究中心（NHGRI）提供了 Genome Hub 和 Ensembl，美国国立生物技术信息中心（NCBI）提供了 Human Genome Central。（见下文）

美国国家人类基因组研究中心（NHGRI）是美国国家卫生研究院（NIH）的一个部门，其 Genome Hub 的网址是 http://www.genome.gov/（然后访问 "online resources"），这个站点给出许多基本的链接。Human Genome Central 由 Ensembl project（见下文）和 NCBI 联合维护，网址是 http://www.ensembl.org/genome/central/和 http://www.ncbi.nlm.nih.gov/genome/central/。Oak Ridge 国家实验室（ORNL）也提供一个人类基因组网站，包含了一些经常被提出的基本问题的资料，地址是 http://www.ornl.gov/hgmis/faq/faqsl.html。

17.2 人类基因组计划的主要结论

作为简介的开始，我们总结了人类基因组计划的主要发现。

① 在人类基因组中经预测的编码蛋白基因大约有 30000～40000 个。这是个令人惊讶的估计，因为我们的基因数与拟南芥（*Arabidopsis thaliana*，26000 个基因）和河豚（puffer fish，33000 个基因）这些更简单的生物的基因数大致相同。

这些结论是从许多资料中总结整理的，包括 IHGSC（2001），Venter 等（2001），David Baltimore（2001），Eric Green 和 Aravinda Chakravarti（2001）的文章，以及 Wellcome Trust Sanger Institute 的网页上所列的"人类基因组的 10 个现象"（http://www.sanger.ac.uk/HGP/publication 2001/facts.shtml）。

② 人类蛋白质组比无脊椎动物蛋白质组复杂得多。脊椎动物具有更为复杂的蛋白模体体系结构的组合。另外，人类基因组通过可变剪切在 mRNA 转录子加工过程中也表现出了巨大的复杂性。

③ 根据最初的报告（IHGSC，2001；Ponting，2001），数百个人类基因是通过水平基因转移从细菌中获得的。随后 Salzberg 等（2001）提出修正，估计经历水平基因转移的有 40 个基因。这些基因与细菌的序列同源，但是在其他脊椎动物和无脊椎动物中似乎缺乏直系同源的基因。

④ 人类基因组 98%以上的部分不编码基因。这些非编码的基因组区域大部分被重复的 DNA 元件所占据，如长散布序列（LINEs，20%）、短散布序列（SINEs，13%）、长末端重复（LTR）转录转座子（8%）和 DNA 转座子（3%）。因此，人类基因组的一半来源于可转座元件。尽管如此，在人类谱系中这些元件的活性已有所下降，但小鼠基因组仍表现出可转座元件的持续的强劲活力。

⑤ 片段扩增在人类基因组中是频繁发生的事件，尤其是在中心体周围（pericentromeric）和端粒下区（subtelomeric）。人类基因组上的片段扩增比酵母、果蝇或是线虫基因组更为普遍。扩增（duplication）的发生主要有三种方式（Green 和 Chakravarti，2001）：一是串联扩增（tandam duplication），即一段序列在局部区域内的多次复制；二是由处理后的 mRNA 通过逆转录转座引起扩增；三是大片段扩增（segmental duplication），指基因组的某个染色体中完整的大片段向另一个区域的扩增性转移。

在下文中我们会详述这些重复元件。

⑥ 人类基因组中有几十万个 *Alu* 重复，曾被认为是偶然杂乱复制的典型元件。但是，这些元件的分布并不是随机的：它们保留在 GC 含量高的区域，可能对基因组有某些益处。

⑦ 男性减数分裂的突变率大约是女性的两倍，这暗示大部分突变是在男性中发生的。

NCBI 的 SNP 数据库目前列出了超过 400 万的 RefSNPs（build 115 版本，June 2003，网址在 http://www.ncbi.nlm.nih.gov/SNP/snp summary.cgi），详见第 18 章。

⑧ 超过 140 万的单核苷酸多态性（SNP）已被鉴定。SNP 是单个核苷酸的变化，约每 100～300 个碱基对会发生一次。

17.3 通过三个门户网站可获得人类基因组数据

人类基因组测序由两个不同的组织分别完成。一个是公共的组织——国际人类基因组测

塞莱拉公司的网址是 http://www.celera.com/。

国际人类基因组测序计划的领导者是 Francis Collins，他是 NHGRI 的主管。可以从 NHGRI 的网址获得其任务信息：http://www.genome.gov/。

序联盟 2000 年 6 月绘出基因组序列草图，并于 2001 年 2 月发表（IHGSC，2001）。其发表的同一周，私人基金支持的塞莱拉公司也发布了他们的测序结果（Venter 等，2001）。由于塞莱拉的结果是私人的，大部分研究人员无法获得他们的序列、组装与注释信息，因此我们也不详加介绍。

公共的测序计划对人类基因组序列进行了多次独立的组装和注释（表 17-1）。组装在 David Haussler 的领导下由 UCSC 协同完成，Greg Schuler 领导的研究组在 NCBI 也进行了组装。注释由 Ewan Birney 主持的 Ensembl project 完成。NCBI 与 UCSC 的 Jim Kent 和 David Hussler 分别对注释进行了分选。

表 17-1 人类基因组公共数据的主要站点

资 源	简 述	URL 网址
Ensembl	EBI 和 Wellcome Trust 研究所的数据	http://www.ensembl.org
National Center for Biotechnology Information	染色体、各种图谱和位点的查看	http://ncbi.nlm.nih.gov/genome/guide
University of California at Santa Cruz	含基因组序列草图的组装	http://genome.ucsc.edu/

17.3.1 NCBI

Whitehead 研究所/MIT 基因组研究中心的 Nat Goodman（2002）对这些资源进行了相互比较和讨论。

NCBI 的人类基因组资源在 http://www.ncbi.nlm.nih.gov/About/Doc/hs_genom-eintro.html。

通过 NCBI 获取人类基因组数据有两个主要的途径。可以从 NCBI 的主页上选择人类基因组资源 "human genome resources"，这个选项提供了到各个染色体资料和一系列网络资源的链接。另外也可以选择人类图谱查看页面 "human map viewer"，这是 Entrez 基因组检索系统的一部分（图 17-1），在页面中单击一条染色体或输入查询文本可以搜索需要的信息。

人类图谱页面整合了人类基因序列和细胞遗传图谱、遗传连锁图谱、放射性杂交图谱以及酵母人工染色体（YAC）。例如输入搜索词 "rbp4"（图 17-1），就可链接到图谱页面［图 17-2(a)］。这个图显示了多种多样的信息，其中一些可选项见图 17-2(a)（箭头 6）、图 17-2(b) 和表 17-2。一个典型的查询结果图（图 17-1）包括指向相关资源（如人类和小鼠 Unigene）的各种链接，一份人类/小鼠同源比较图，以及到包含一些存疑的基因的克隆重叠群的链接。每一幅图都有描述各基因模型可信程度的置信代码 "evidence codes"（表 17-3），同样也有标明特定基因存在的证据信息的查看页面 "evidence viewer"（图 17-3），突出了核酸序列可能偏差、外显子与内含子的边界及其他方面的注释。在图 17-3 的例子中，页面上就显示了根据基因 *RBP4* 每一个预测的外显子确定的表达序列标签（ESTs）的密度。

17.3.2 Ensembl

直接访问人类基因图谱页面：http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map search。

Ensembl 是一个有关人类基因组和若干其他物种基因组的全面的资源（Hubbard 等，2002）。它能有效地与广泛的基因组工具进行链接，这些工具集中用于已知和预测基因的注释。除了使基因的

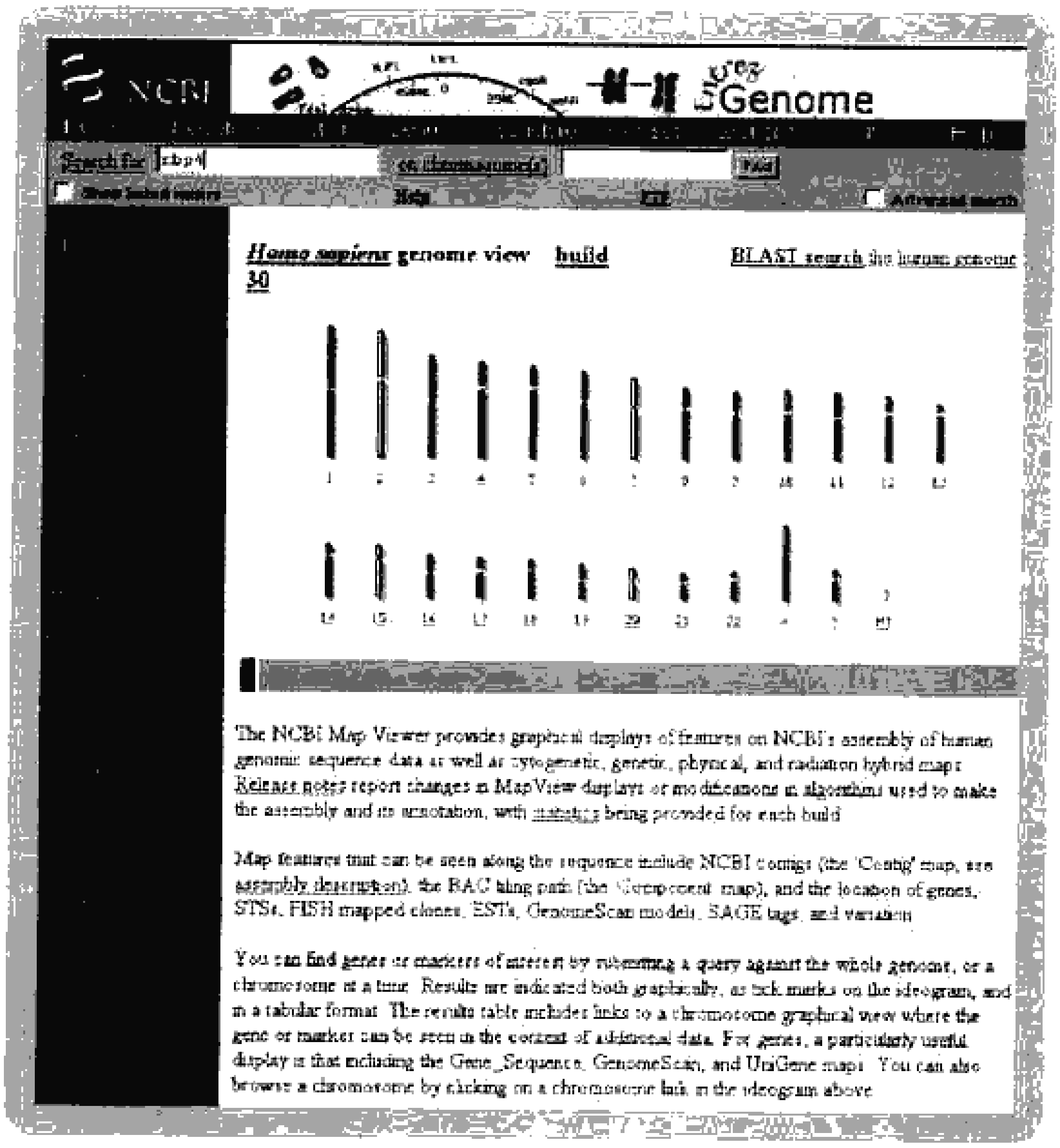


图 17-1 从 NCBI 主页可链接到人类图谱页面，包含了人类基因组序列的细胞、遗传、物理和放射性杂交图谱。图中输入了检索词“rbp4”，结果链接到 10 号染色体（见图 17-2）

注释信息更易获得外，Ensembl 还提供了支持各种模式基因预测的数据，详见下文。关于 Ensembl 目前的人类信息的统计数据见表 17-4。

> Ensembl 的网址在 http://www.ensembl.org。人类数据库在 http://www.ensembl.org/Homo_sapiens/。我们在第 16 章叙述了小鼠、大鼠、斑马鱼、河豚和蚊子的 Ensembl 计划。

表 17-2 NCBI 图谱页面的显示选项

细胞遗传图	Marshfield 图	序列图	SAGE 标签
图标符号	放射性杂交图谱	克隆	STS
荧光原位杂交(FISH)克隆	GeneMap99-G3	重叠群	转录子(RNA)
基因-细胞遗传	GeneMap99-GB4	组成	Hs_UniGene
Mitelman 断裂点	NCBI RH	CpG 岛	Mm_UniGene
病变/疾病	Stanford-G3	DbSNP 单型	多样性
遗传图谱	TNG	GenBank_DNA	
解析	Whitehead-RH	基因	
Genethon 图	Whitehead-YAC	基因组扫描	

表 17-3 NCBI 图谱页面基因模型的置信代码

基因颜色	置信代码	用以建立基因模型的根据
蓝色	C	确定的基因模式——基于 mRNA 和表达序列标签(ESTs)或 mRNA 比对
亮绿色	E	仅依据 EST 数据
暗褐色	PE	GenomeScan 的预测加上 EST 数据
亮褐色	P	仅依赖于 GenomeScan 预测
橘黄色	?	矛盾：基因模型和 mRNA 序列存在偏差
	I	临时的 LocusID——基于 mRNA 或加上 EST 的数据的基因组比对，但用以比对的转录子还没有一个清楚的 LocusID

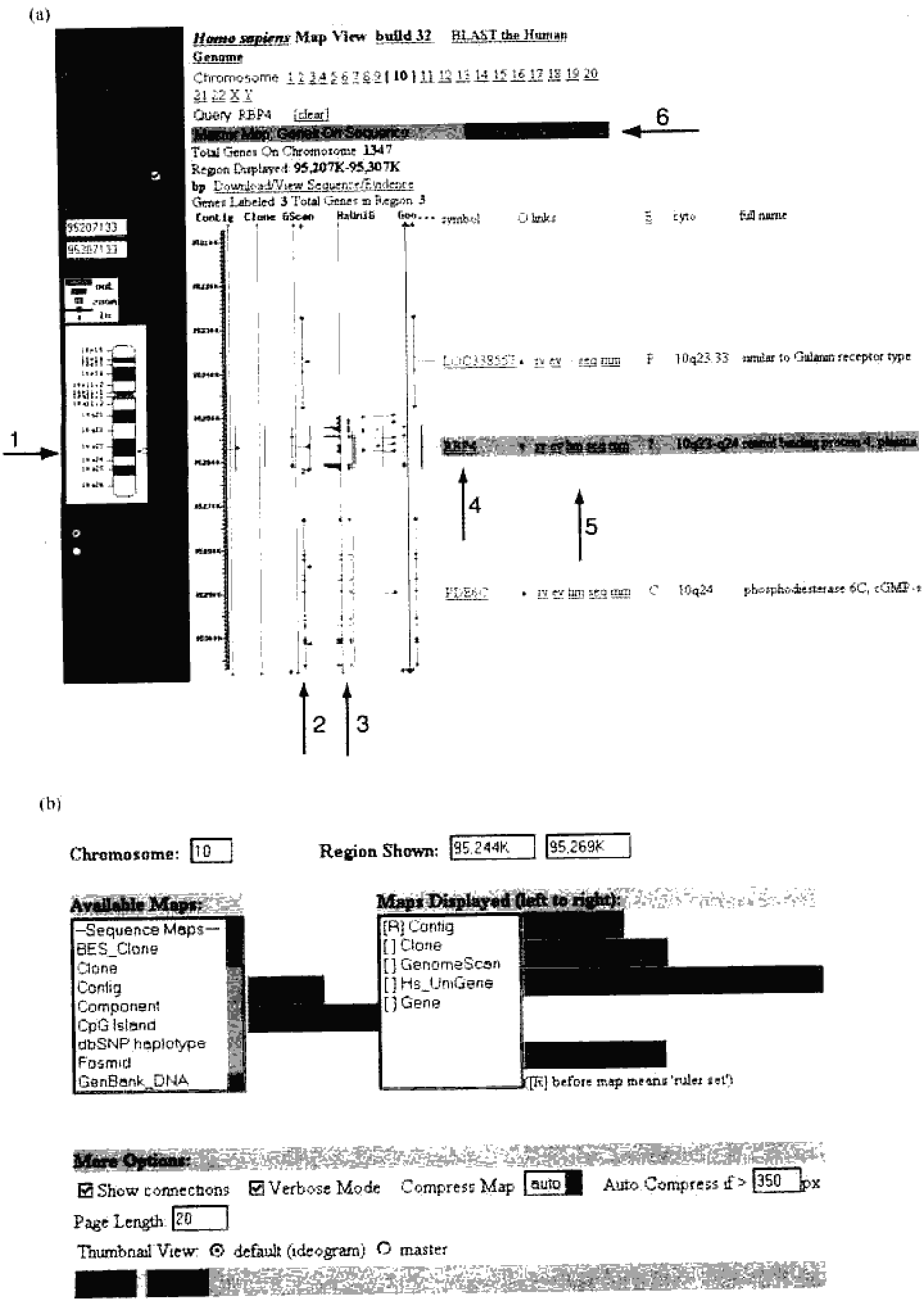


图 17-2 (a) NCBI 的人类图谱页面显示出 10 号染色体包含 *RBP4* 基因的区域。位点用一个图形符号指出（箭头 1），上面有一个缩放工具。主要的信息分栏显示，从左到右分别是细胞遗传图（箭头 2）、*RBP4* 在 UniGene 的链接（箭头 3）、可对基因进行 10kbp～10Mbp 缩放的物理图和一个指向 LocusLink 的链接（箭头 4）。其他的链接（箭头 5）还有 sv（Entrez 序列的查看）、ev（证据信息的查看）、hm、seq 和 mm。在页面的顶端，点击显示设置（箭头 6）可出现一系列的下拉菜单。(b) 下拉菜单：这些选项可以改变显示在主图中的字段

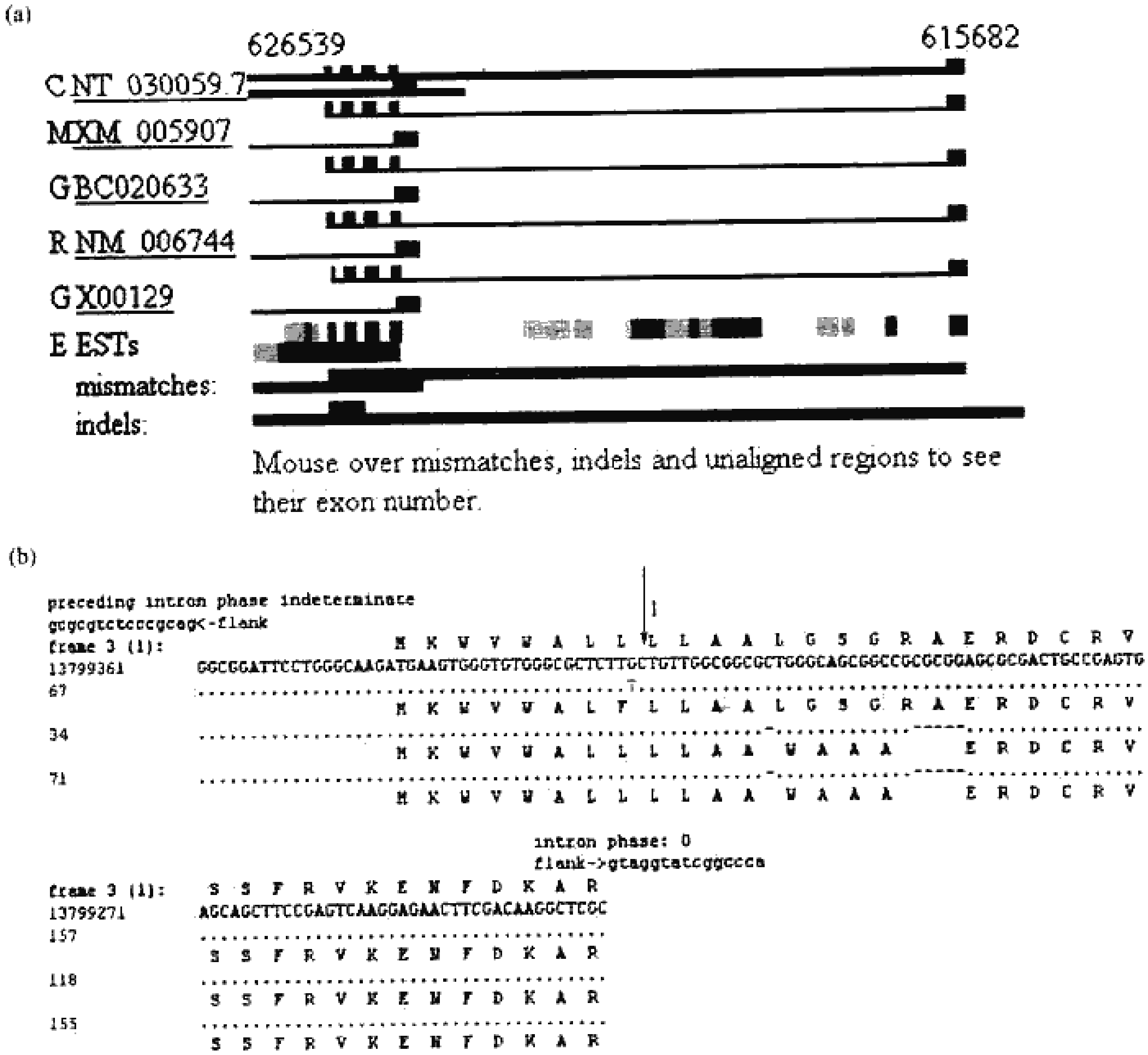


图 17-3 (a) NCBI 证据信息页面提供了人类基因的结构信息，也有指向基因组重叠群(C)、RefSeq 的 mRNA 模式 (M)、GenBank mRNA (G)、RefSeq mRNA (R) 以及表达序列标签 (ESTs) (E) 的链接，EST 链接用不同颜色标明各基因位点上 EST 的密度。此外该页面还记述了每个外显子的实验证据。(b) 例如，如在 *RBP4* 基因的第 2 个外显子上检测到了一个核苷酸的错配 (箭头 1)，则其外显子内含子边界的标注就比较可疑

表 17-4 Ensembl 人类信息网络服务的统计数据 (2003 年 6 月)

Ensembl 预测基因	24847	重叠群	33840
Genscan 预测基因	58770	克隆	26759
Ensembl 基因外显子	248449	碱基对	3242415757
Ensembl 转录子	37347		

注：见 http://www.ensembl.org/Homo_sapiens/stats/。

> 图谱页面有指向 LocusLink 的链接，而如果从 LocusLink 进入的话 (图 2-8)，也可以点击 "map" 直接链接到图谱页面。

Ensembl 的主页上可以输入检索文本 (例如 RBP4)，可以进行 BLAST 搜索，也可以根据染色体进行浏览 (图 17-4)。可以由 6 个主要的入口访问 Ensembl 数据库。

① 基因视图 (Gene view) 允许检索词搜索 (如 RBP4)，并得到典型的 Ensembl 基因报告 (图 17-5 和图 17-6)。这个报告包括到重叠群视图的链接 (图 17-7)、"transview" 置信报告 (图 17-8)、蛋白报

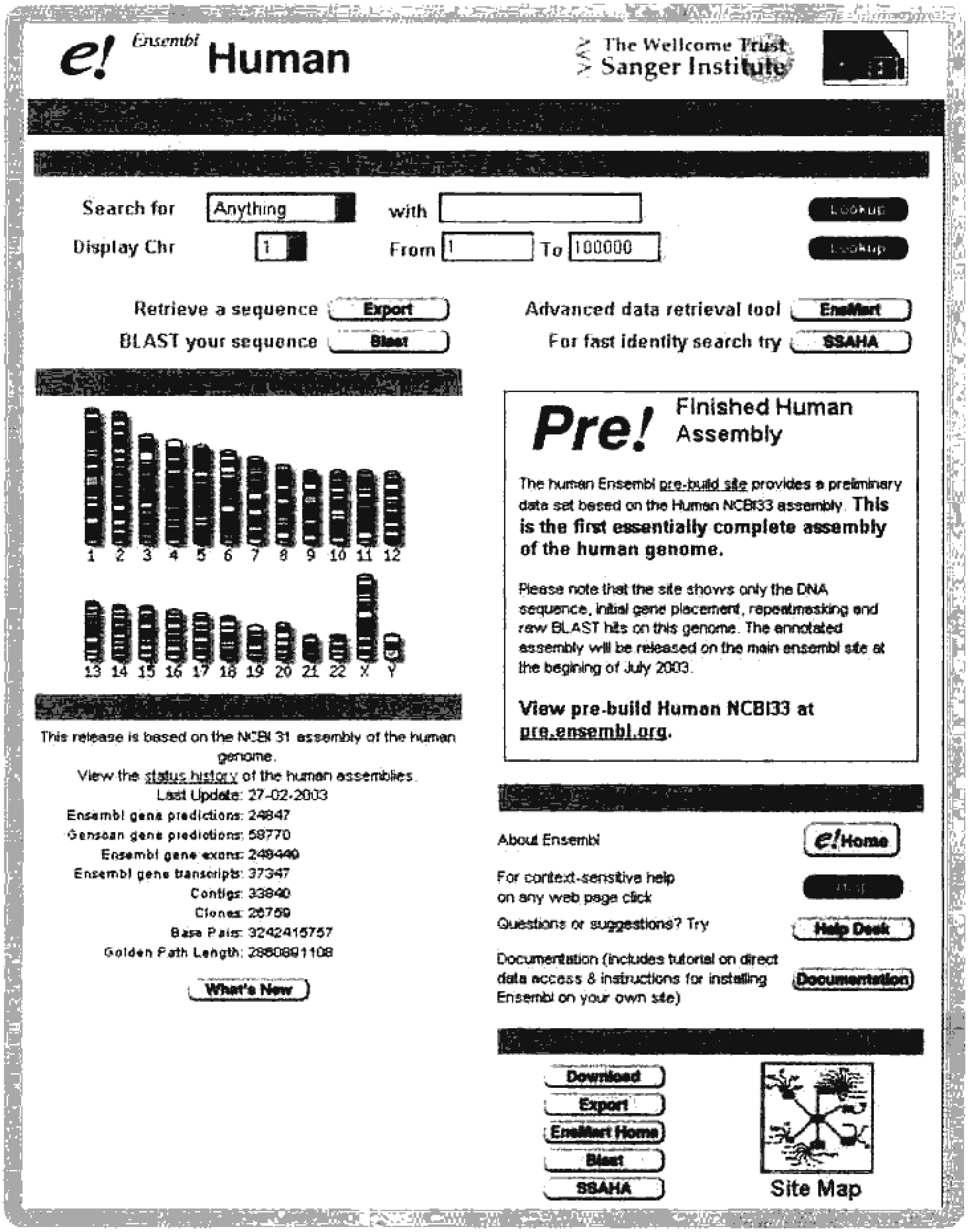


图 17-4 Ensembl 人类基因组的主页（http://www.ensembl.org/Home_sapiens/）。输入检索词，如 RBP4（页面顶部），就可以开始搜索。检索字段可以从下拉菜单中选择（如疾病、结构域、EST、家族、基因、mRNA、SNP 等）。通过单击染色体图示或 BLAST 服务也可以进行搜索

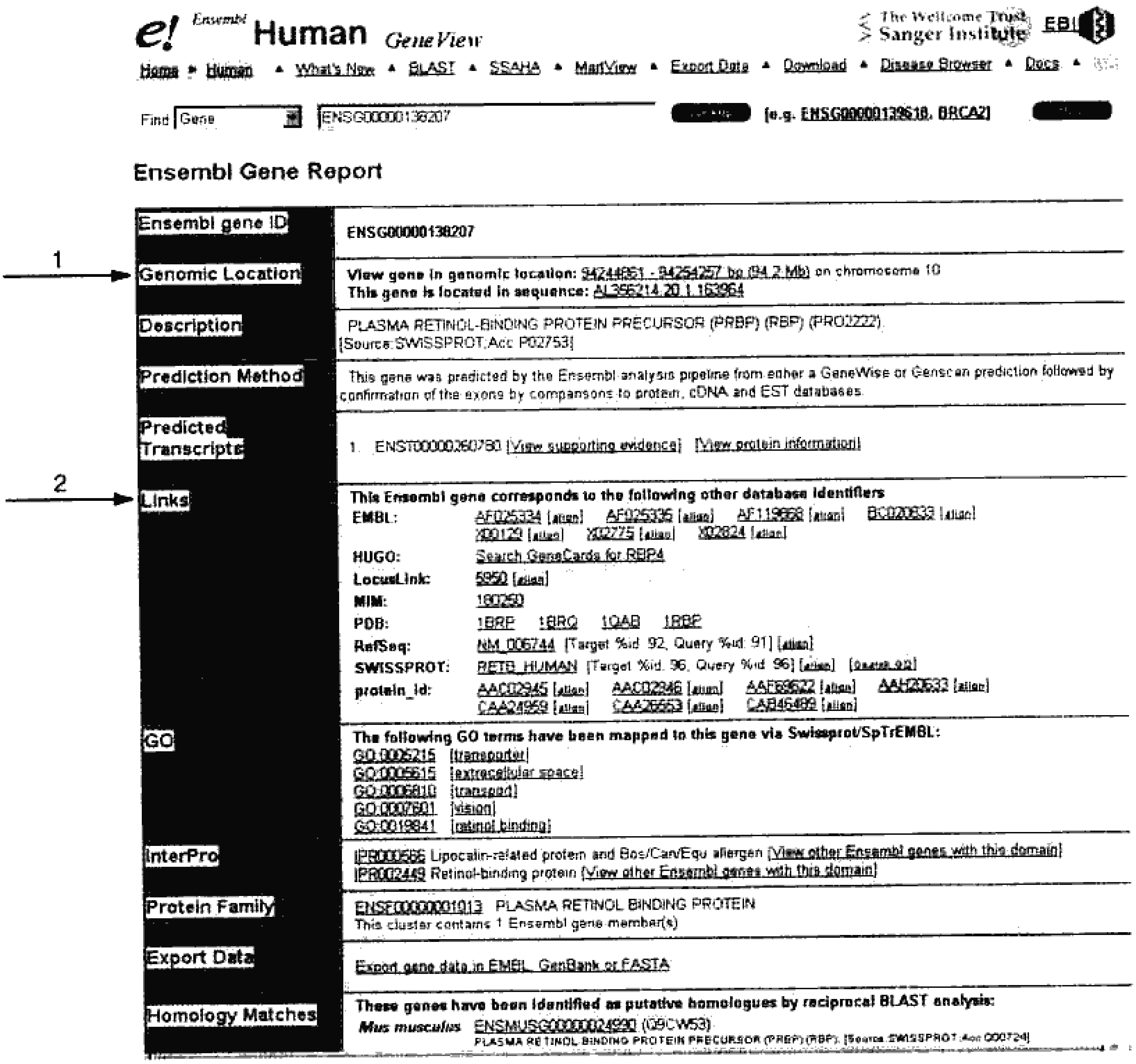


图 17-5 典型的 Ensembl 基因报告（本图仅示上半部分）。本图是人类 *RBP4* 的报告，包括指向基因组重叠群的链接（箭头 1；图 17-7）和我们已经学习过的几个数据库的链接（箭头 2）

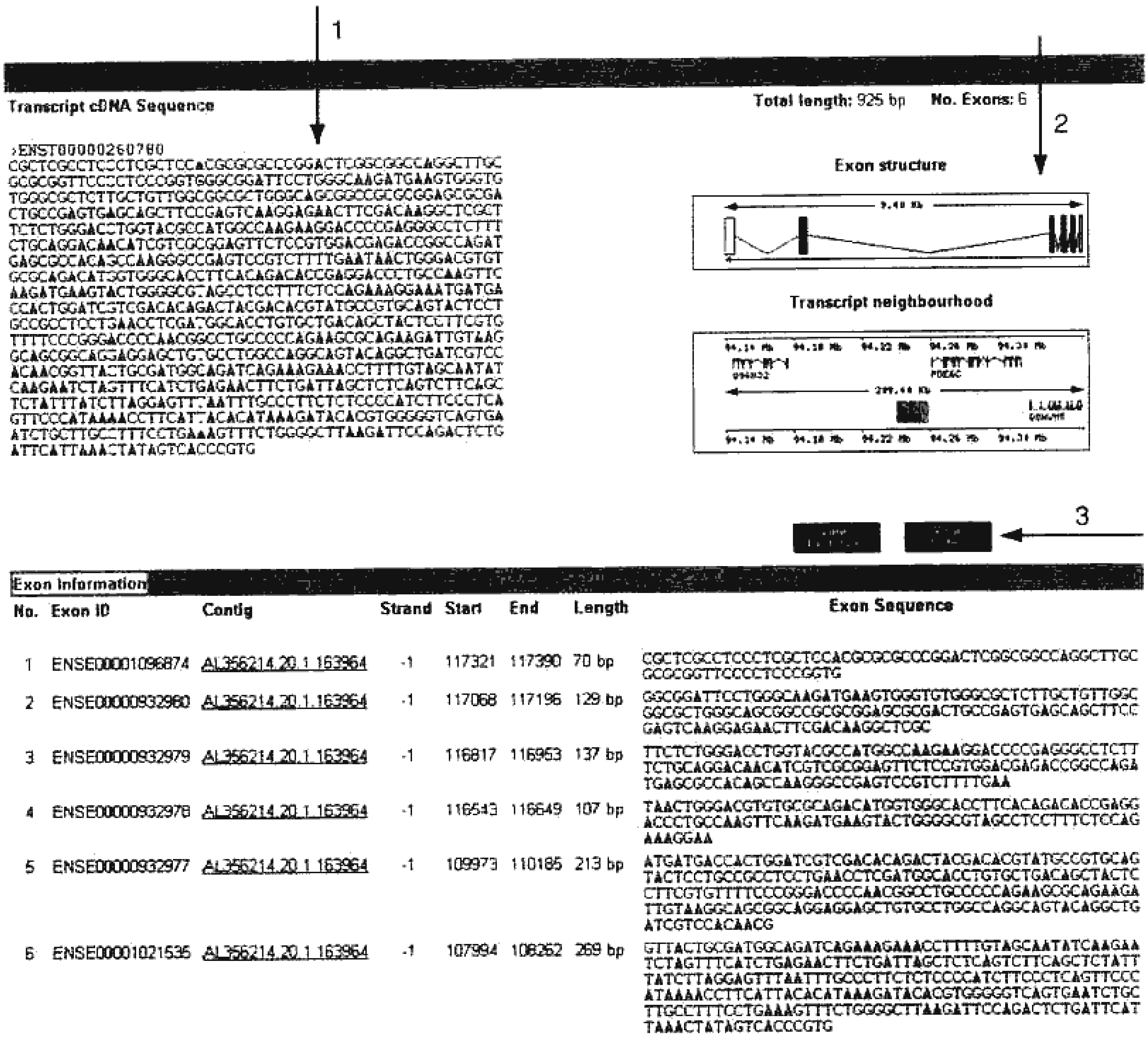


图 17-6 人类 *RBP4* 基因的报告的中间部分给出了 *RBP4* 的 cDNA 序列（箭头 1）和外显子的分布（箭头 2）以及相邻的基因。"view evidence" 和 "view protein" 链接（箭头 3）的数据分别见图 17-8 和图 17-9。报告的最后包括外显子和预测的剪切位点（本图未显示）

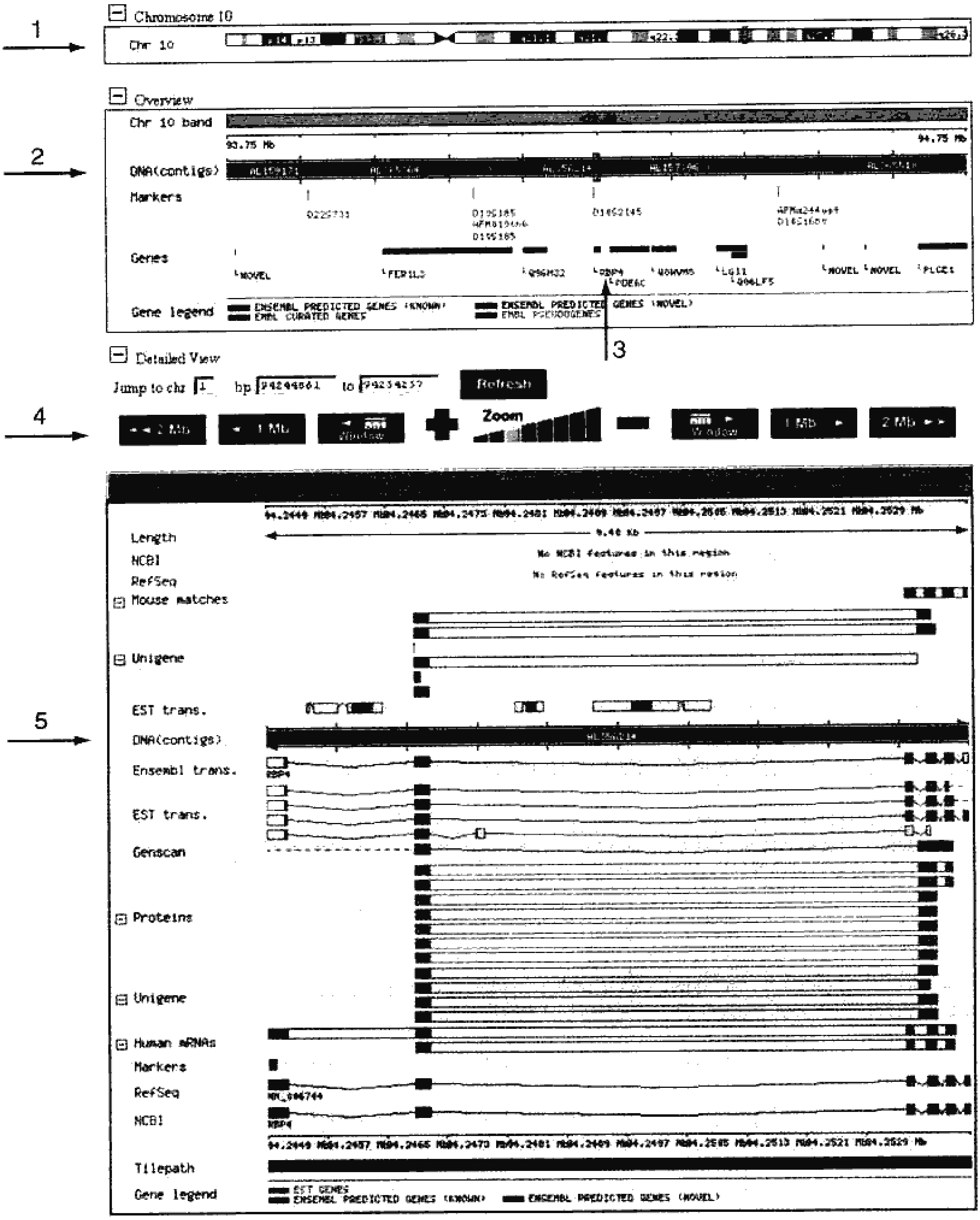


图 17-7 Ensembl 重叠群视图（从 *RBP4* 的基因报告链接而得，见图 17-5）给出第 10 号染色体的图示（箭头 1），其中包含 *RBP4* 基因的位点（方框所在）。Overview 部分有 10q23.33 区域的重叠群的概貌（箭头 2）和含有 *RBP4* 的注释基因（箭头 3）。Detail view 部分可以滚动和缩放（箭头 4），包括一个重叠群的示意图（箭头 5）及其两条链的注释（含链接到 EST 的 *RBP4*、标记序列和预测转录子的链接）

e! Ensembl Human *TransView* — The Wellcome Trust Sanger Institute EBI

Home ▸ Human ▴ What's New ▴ BLAST ▴ SSAHA ▴ MartView ▴ Export Data ▴ Download ▴ Disease Browser ▴ Docs ▴

Find All | ENST00000260780 | Lookup | [e.g. ENST00000157775] | Help

Transcript ENST00000260780

No.	Exon	Start	End	Strand	Gene ID	Contig ID
1	ENSE00001096874	117321	117390	-1	ENSG00000138207	AL356214.20.1.163964
2	ENSE00000932980	117068	117196	-1	ENSG00000138207	AL356214.20.1.163964
3	ENSE00000932979	116817	116953	-1	ENSG00000138207	AL356214.20.1.163964
4	ENSE00000932978	116543	116649	-1	ENSG00000138207	AL356214.20.1.163964
5	ENSE00000932977	109973	110185	-1	ENSG00000138207	AL356214.20.1.163964
6	ENSE00001021535	107994	108262	-1	ENSG00000138207	AL356214.20.1.163964

Supporting evidence for exons

Below is a table of database hits having overlaps with each exon in the transcript. The database hits are the results of a series of blast runs against genscan predicted peptides. They are ordered by:

- Data library,
- Top scoring, exon vs sequence, hit
- Number of exons.

A threshold has been applied to the supporting evidence table.

- For Blast hits this value is 80

Note:

- An exon without supporting evidence means that it was generated by aligning a protein to the genomic sequence using genewise.
- Low scoring evidence is greyed out

Data Library	Sequence	Definition	Exons 1	2	3	4	5	6
	NP_006735	gi|5803139|ref|NP_006735.1| (NM_006744) retinol-binding protein 4, plasma precursor; retinol-binding protein 4, plasma; retinol-binding protein 4, interstitial [Homo sapiens]						■
	RETB_HUMAN	RETB_HUMAN P02753 O43478 O43479 Q8WWA3 Q9P178 CAA24959.1 AAH20633.1 CAA26553.1 CAB46488.1 AAF69622.1 AAC02945.1 AAC02946.1 Desc: Plasma retinol-binding protein precursor (PRBP) (RBP) (PRO2222).						■
NULL	[illegible] [illegible] [illegible] NONE		■					
dbEST	AW027400.1	gi|5886156|gb|AW027400.1|AW027400 wv73c02.x1 Soares_thymus_NHFTh Homo sapiens cDNA clone IMAGE:2535170 3' similar to gb:X00129 PLASMA RETINOL-BINDING PROTEIN PRECURSOR (HUMAN);			■		■	
dbEST	AI132990.1	gi|6360306|gb|AI132990.1|AI132990 HA1576 Human fetal liver cDNA library Homo sapiens cDNA			■		■	
dbEST	BI599359.1	gi|15492298|gb|BI599359.1|BI599359 603246016F1 NIH_MGC_96 Homo sapiens cDNA clone IMAGE:5288686 5'		■	■		■	

图 17-8 Ensembl 基因报告（图 17-6，箭头 3）链接到特定的转录子的置信视图。该图包括 *RBP4* 六个外显子每一个的绘制根据。各个外显子在各数据库中的情况都列成表（本图只列出一小部分）

告（图 17-9）、数据库链接（如 RefSeq、Swiss、Port、LocusLink 和 Interpro）以及 Ensembl 同源匹配。基因视图还提供转录 DNA 序列和外显子内含子边界信息（剪切信息）。其中，置信报告的链接对于评价基因结构很有帮助，因为它提供了实验数据信息。

② 重叠群视图（Contig view）用于对整个染色体的查找，但也可以查看一个小区域的详细情况。图 17-7 演示了一个例子。重叠群视图整合了不同数据来源（如 UniGene、RefSeq）的各种扩展信息。“TilePath”显示了用于当前序列组装的 BAC 克隆。

③ 定点视图（Anchor view）可以让用户在染色体上选两个特征位点作为锚定点（anchor points），然后展示这两点之间的区域。

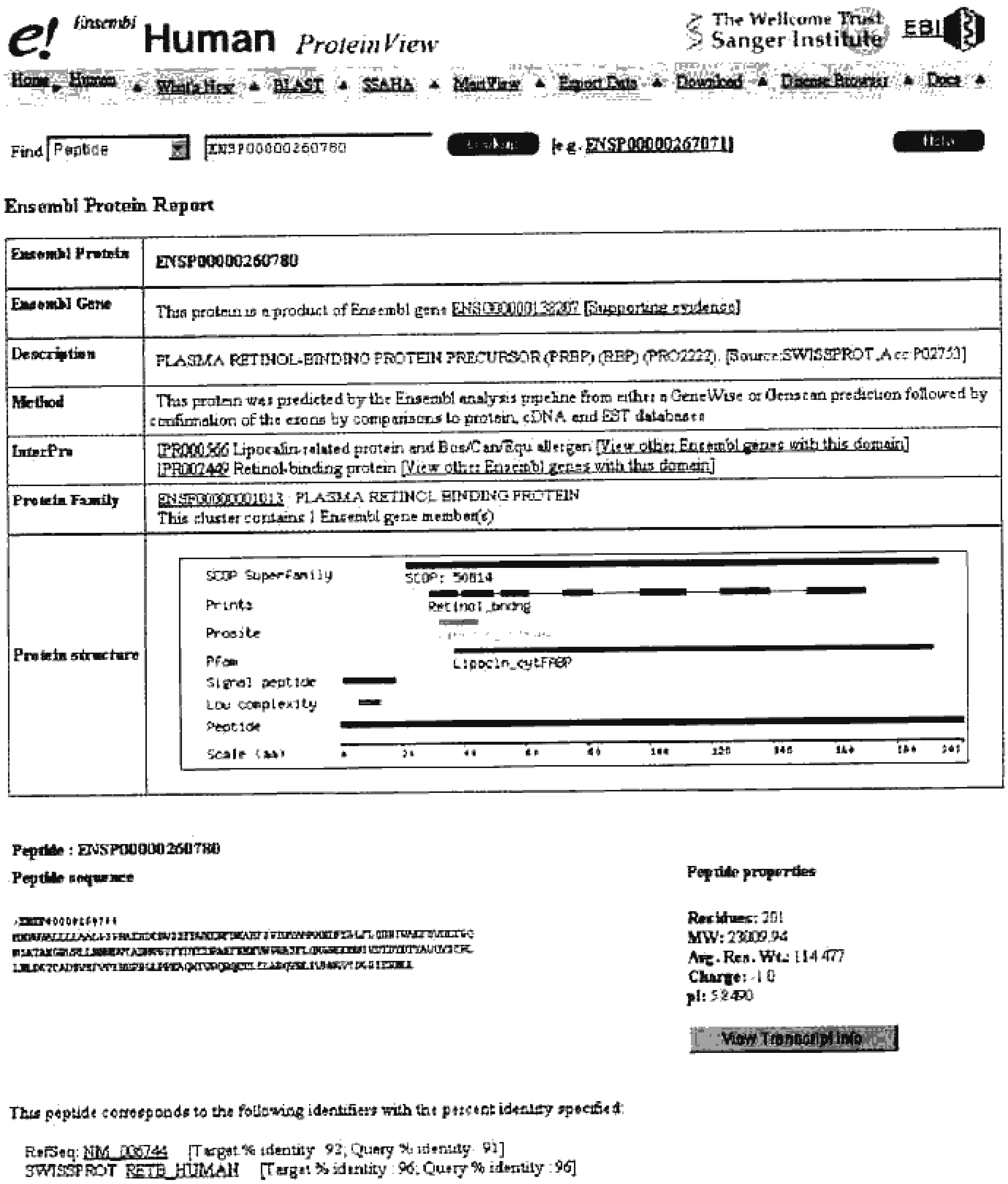


图 17-9 *RBP4* 的 Ensembl 蛋白质报告（图 17-6，箭头 3）显示了 *RBP4* 编码蛋白的各方面特征，如氨基酸序列、结构预测，以及它在 PRINTS、Pfam 和其他蛋白数据库中的结构域信息

④ 疾病视图（Disease view）链接到 OMIM（第 18 章）的疾病数据。以 *RBP4* 为例，该视图选项链接到相关的 *RBP4* 缺损的 OMIM 病例。

⑤ 图谱（Map view）给出染色体图示，包括已知基因、GC 百分率和 SNP（图 17-10）。通过点击物种的比较链接（图 17-10，箭头 1），还可以看到 *RBP4* 也定位于小鼠第 19 号染色体上的相应区域（图 17-11）。

⑥ 细胞视图（Cyto view）显示基因、BAC 末端克隆、重复元件和基因组 DNA 区域的"tiling path"。

Ensembl 的网址在 http://www.ensembl.org。人类数据库在 http://www.ensembl.org/Homo_sapiens/。我们在第 16 章叙述了小鼠、大鼠、河豚鱼、斑马鱼和蚊子的 Ensembl 计划。

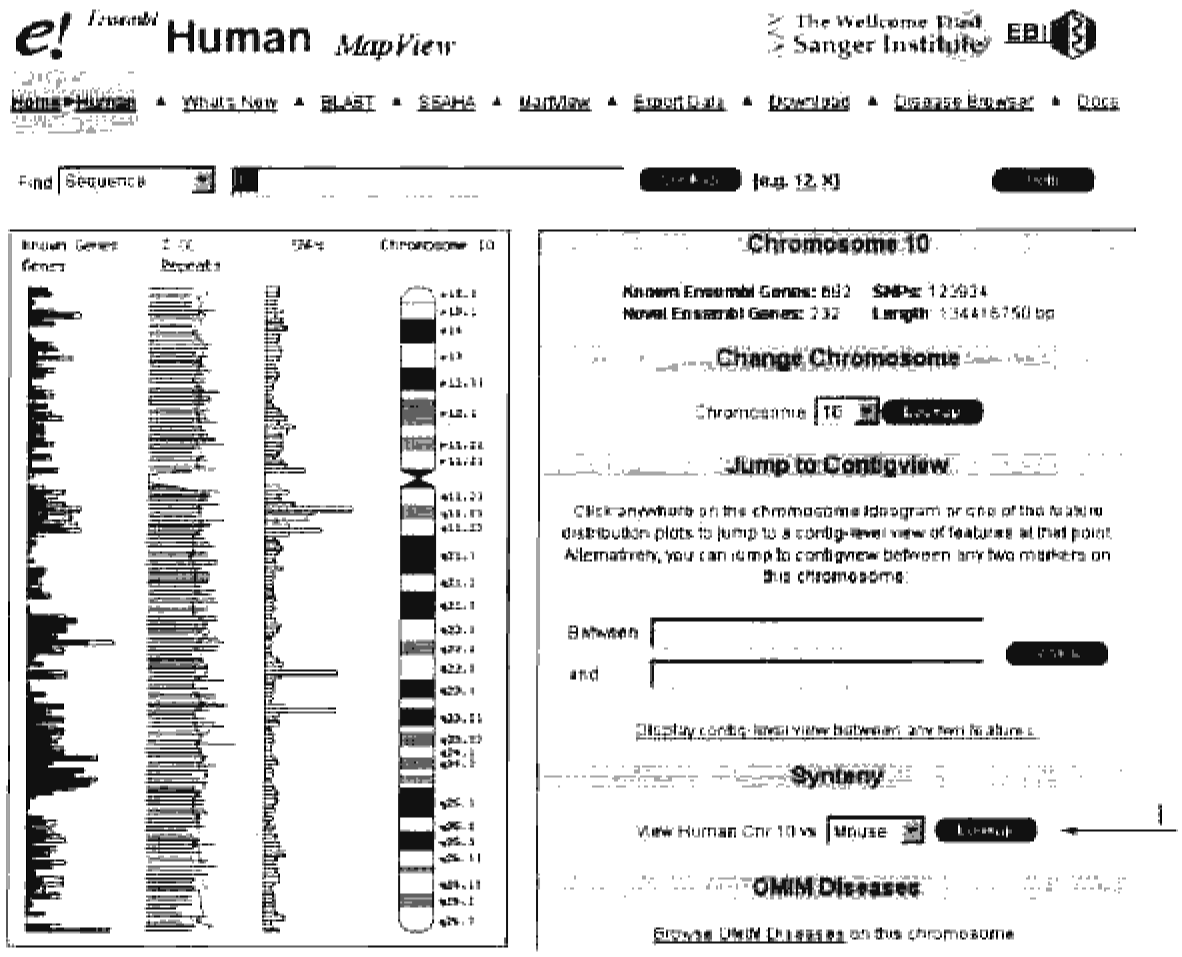


图 17-10 Ensembl 染色体图谱可显示每个染色体及其基因、SNP、GC 的分布等相关数据，还有一个到小鼠基因组相应位置的链接（箭头 1）（见图 17-11）

17.3.3 UCSC 人类基因组浏览器

关于 Ensembl 的 BLAST 服务有一个例子，见图 5-1 和图 5-2。

UCSC 基因组生物信息学网站为 http://genome.ucsc.edu/index.html，是由 David Haussler 的小组开发的（Kent 等，2002）。

UCSC 网站的数据基于 NCBI 组装的人类基因组。

在 UCSC（university of California at Santa Cruz）的人类基因组注释被称为"Golden Path"。除了 Ensembl 和 NCBI 以外，UCSC 人类基因组浏览器也是获取人类和其他脊椎动物基因组信息的主要网络资源。

UCSC 浏览器支持关键词、基因名或其他文本查询，也可以用类似 BLAST 的比对工具 BLAT（Kent，2002）（见第 5 章）来进行数据检索。图 17-12 是查询视黄醇结合蛋白 rbp4 的输出结果：顶部可以调节 DNA 的大小，窗口可以沿染色体拖动；主窗口是 *RBP4* 基因的注释行；窗口底部有许多可供选择的其他注释行。注释行包括以下几类（Kent 等，2002）。

① 重叠群和缺口的组装。

② mRNA 和 EST 的比对以及从 SAGE 或微阵列获得的基因表达数据。

③ 多种方法的基因预测，如 *RBP4*：不同方法对外显子的预测结果有时会不一致。

④ 跨物种的同源性。

⑤ SNPs。

⑥ DNA 重复片段。

⑦ 放射性杂交数据。

使用上述扩展的注释行的一个例子见图 17-13。所得数据包括与 *RBP4* 匹配的非人类的 mRNA，以及在 BLAT 中小鼠、黑猩猩、鸡等和 *RBP4* 的比对相匹配的非人类的 mRNA，还有从 60 个细胞系的 8000 个表达基因中搜集的 *RBP4* 的基因表达数据（Ross 等，2000）。

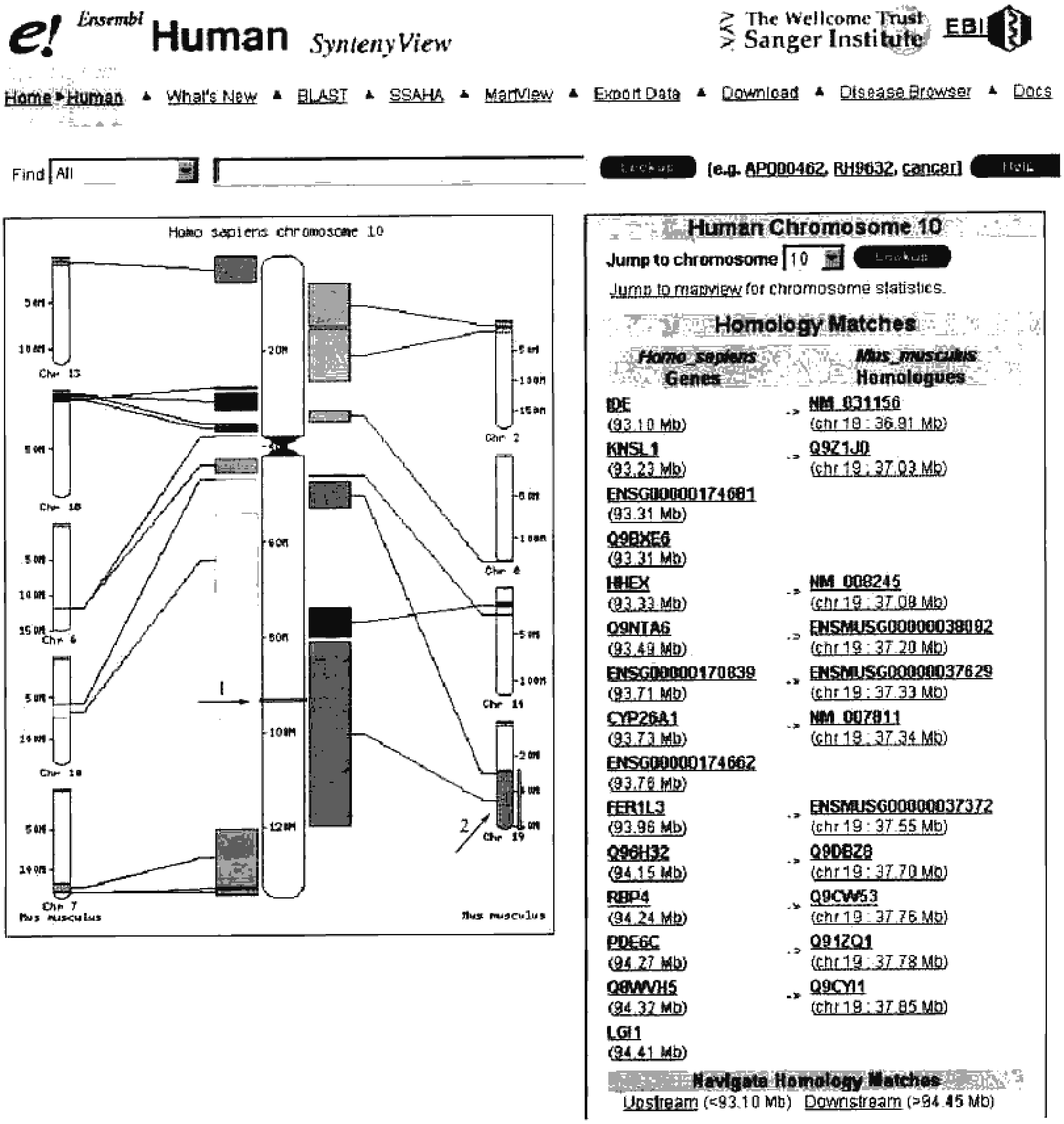


图 17-11　Ensembl 图谱链接到人类/小鼠基因保守区对应位置的比较图。图中显示了人类第 10 号染色体，包括含有 *RBP4* 基因的区域（箭头 1 和右边的基因列表），以及小鼠第 19 号染色体的对应区域（箭头 2）

(a)

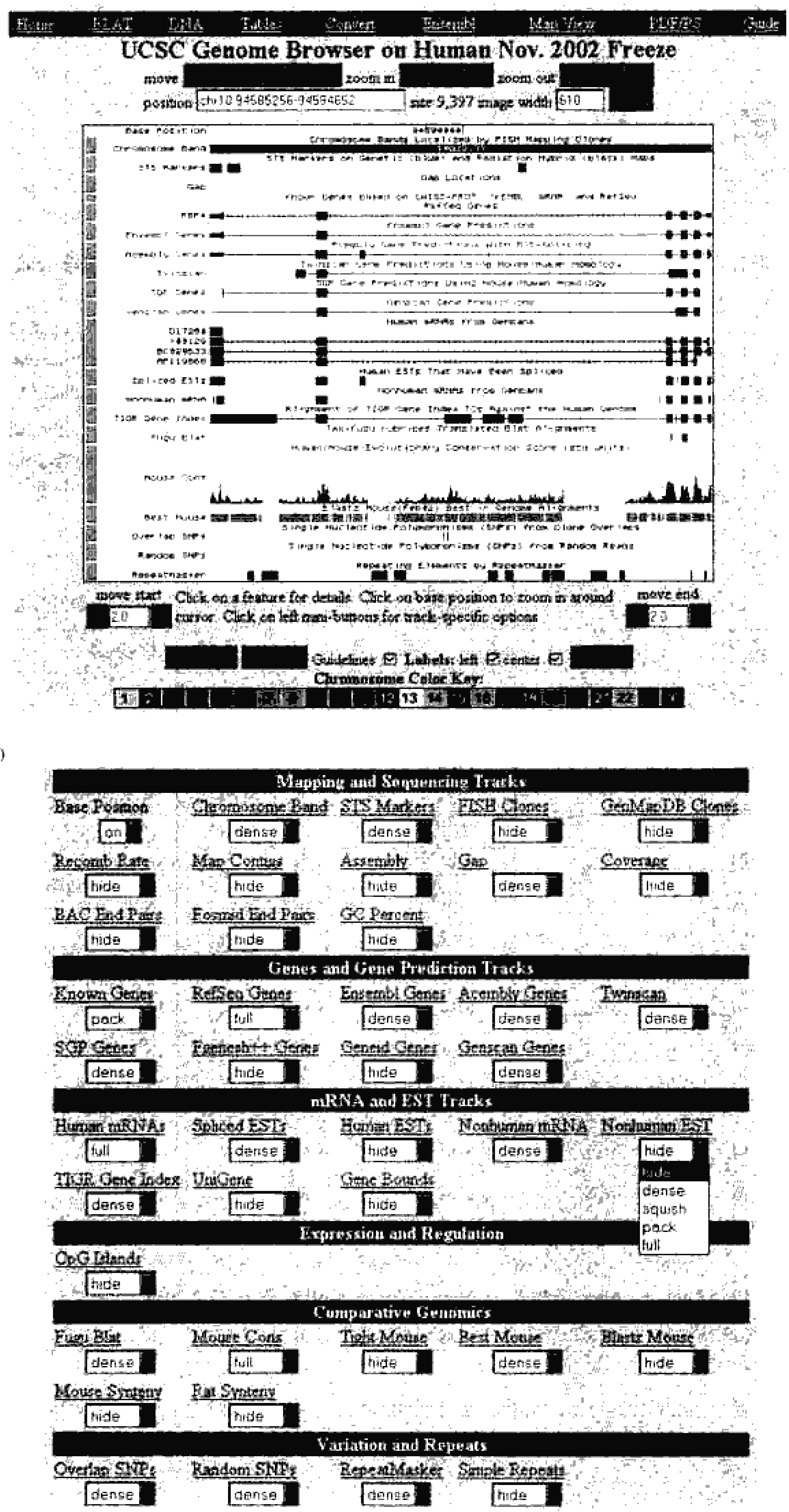


图 17-12 UCSC 的人类基因组浏览器（http://genome.ucsc.edu）（Kent 等，2002）。通过顶部的按钮，窗口可以沿染色体拖动，基因组注释显示在窗口中。窗口底部还有一系列控制选项，它们使各类注释的显示（如重复 DNA、比较基因组和表达数据）可以被动态地调整

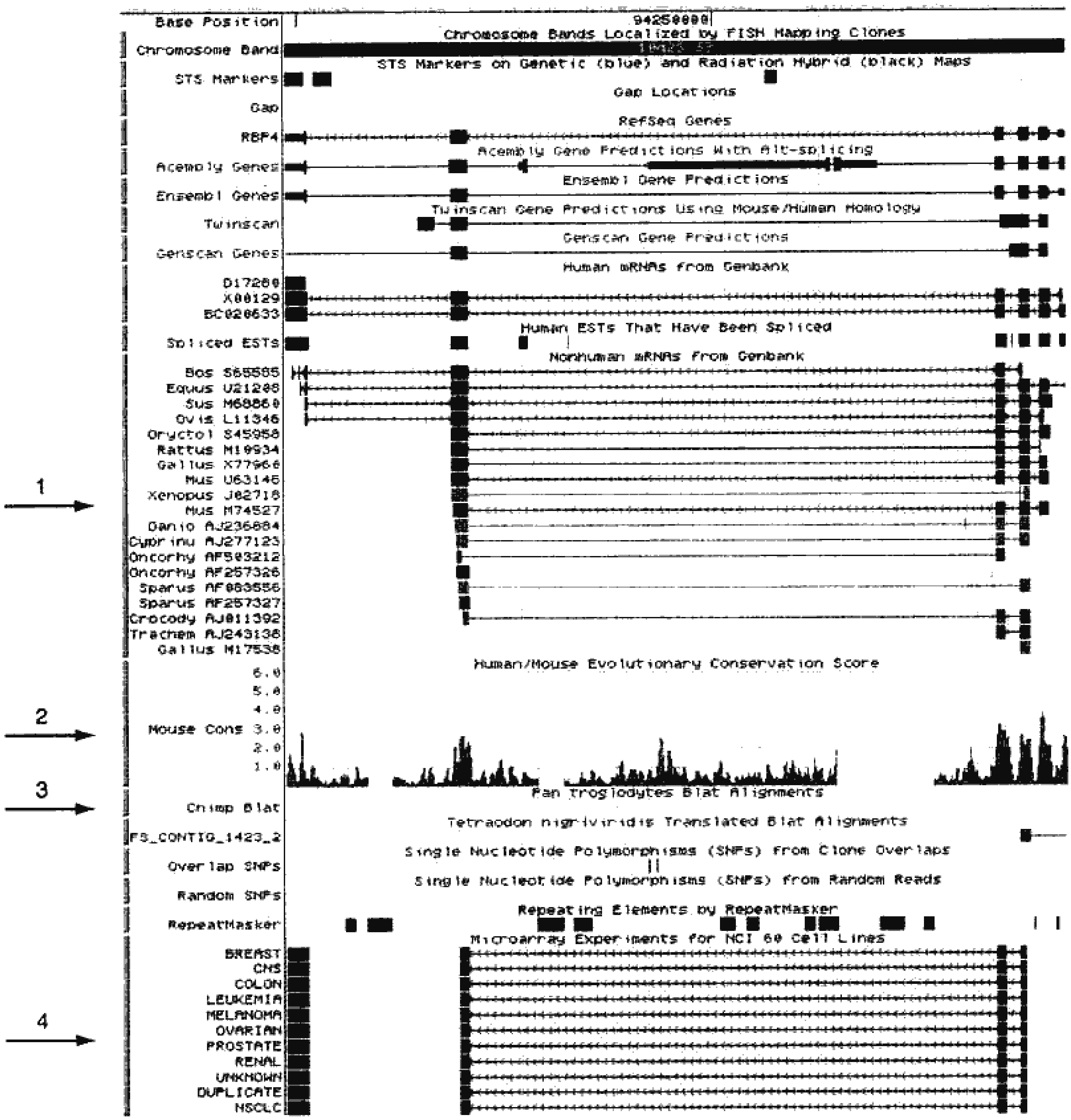


图 17-13 UCSC 的人类基因组浏览器提供易于调整的注释。图中 *RBP4*（图 17-12）的扩展记录给出了外显子和人类以外物种的匹配（箭头 1），并显示了相应的小鼠/人类基因保守区域（箭头 2），还有与黑猩猩基因的匹配（箭头 3），以及测量 60 个细胞系中 8000 个基因表达的"NC160"实验的表达数据（箭头 4）（这表明 *RBP4* 的表达在癌细胞系中上调）

除了提供这些注释外，UCSC 基因组浏览器允许用户为任何数据添加自己的注释。实际上，在网站所有注释中 UCSC 完成的占大约 50%，而另一半就是使用这个网站的研究者们提供的（Kent 等，2002）。

17.4 人类基因组计划

关于人类基因组计划的两篇文章发表于 2001 年 2 月，阐述了对人类基因组的初步了解

(IHGSC，2001；Venter 等，2001)。本章剩下的部分就是根据国际测序联盟的文章（IHGSC，2001）的纲要来叙述的，但不会列举所有的发现，而是集中于一些选定的主题。2001 年报告的序列占人类基因组的 90%，还有其他文章对人类第 20 号、第 21 号和第 22 号染色体做了详细的描述（Dunham 等，1999；Deloukas 等，2001；Hattori 等，2000）。完成的测序精度至少达到了 99.99%，覆盖了常染色质的 95%以上，几乎无缺口，因此达到了人类基因组计划的目标。

常染色质是基因组中含有基因的部分。

17.4.1 人类基因组计划的背景

人类基因组计划最早由美国国家研究委员会提出（1988），旨在建立人类基因组的遗传、物理和序列图谱。与此同时，其他模式生物的测序工作也得到了广泛支持［如细菌、酵母（*Saccharomyces cerevisiae*）、线虫（*Caenorhabditis elegans*）、果蝇（*Drosophila melanogaster*）和小鼠（*Mus musculus*）］。

人类基因组计划的主要目标见表 17-5。

美国科学院出版社（http://www.nap.edu）在网上免费提供 1988 年的这份报告：http://www.nap.edu/books/0309038405/html/。

表 17-5 人类基因组计划的 8 个目标（1998～2003 年）

目标	包含内容
1. 人类 DNA 测序	①2003 年完成人类基因组全部 DNA 测序工作 ②2001 年完成基于图谱的克隆覆盖至少 90%基因组的工作草图 ③保证序列数据完全免费、世界共享
2. 改进与提高测序技术	①使现有技术继续提高效率，并降低成本 ②支持对测序技术进步有重要意义的新技术的研究 ③开发应用和发展新测序技术的有效方法
3. 人类基因组序列多样性	①开发针对 SNP 和其他序列多样性的快速高通量检测与评估方法 ②5 年内识别大部分已知基因编码区的普遍多型性 ③绘制至少有 100000 个标记的 SNP 图谱 ④建立 DNA 样本和细胞系的公共资源
4. 功能基因组学	①生成代表人类和其他模式生物的全长 cDNA 克隆和序列数据集 ②支持研究非编码序列功能的方法学探索 ③开发基因表达的全面分析技术 ④发展全基因组突变技术 ⑤开发大规模蛋白质分析技术
5. 比较基因组学	①完成线虫(*C. elegans*)基因组和果蝇(*Drosophila*)基因组的测序 ②绘制整合小鼠的物理和遗传图谱，完善小鼠 cDNA 库，2008 年完成小鼠基因组测序
6. 伦理、法律及社会问题	①调查人类基因组测序和遗传多样性研究中的各类问题 ②如何在医疗护理和公共卫生工作中安全有效地利用遗传技术和信息 ③基因组和基因-环境相互关系的知识在医疗领域外应用中产生的新问题 ④新的遗传学知识将会如何影响原有的一系列哲学、神学、伦理学观念，探究种族和社会生态因素对于遗传信息的理解、遗传服务的应用以及政策制定的影响
7. 生物信息学和计算生物学	①数据库内容和效用的改进 ②开发数据产生、获取和注释的新工具 ③开发和改进综合功能研究的数据库和分析工具 ④开发和改进序列相似性、多样性的表示和分析工具 ⑤建立相应机制支持优秀共享软件的高效开发
8. 培训专业技术人员	①培养从事基因组研究的科学人才 ②鼓励为基因组研究人员设立学术职位 ③培养具有交叉学科知识的学者，使他们了解基因组和遗传学，并对伦理、法律或社会科学也有一定造诣

注：来源并改编于 http://www.ornl.gov/TechResources/Human_Genome/hg5yp/goal.html。

人类基因组计划的一部分是ELSI研究，即研究人类基因组计划的实施引起的伦理、法律和社会各方面的广泛问题。美国能源部和国立卫生研究院每年把人类基因组计划预算的3%～5%用于ELSI的研究，使其成为世界上最大的生命伦理学计划。

ELSI的网址为http://www.ornl.gov/ghmis/elsi/elsi.html。

ELSI研究的问题包括：

① 谁拥有基因信息？

② 谁有权获得基因信息？

③ 基因组信息对大众的社会生活有什么影响？

④ 新的生殖技术会产生什么样的社会问题？

⑤ 应该如何管理遗传检查，如何判断其合法性和可靠性？

⑥ 应该在多大程度上限制对基因的控制和改变？

⑦ 进行过基因处理的食品是否对人类健康有潜在的危险？

17.4.2 测序战略：用分级鸟枪法得到序列草图

国际人类基因组测序联盟的测序方法是分级鸟枪法（hierarchical shotgun），基本原理如下：

① 鸟枪法测序可以应用于不同大小的DNA分子，包括质粒（几千碱基对）、黏粒克隆（40kb）、酵母，以及BAC（1～2Mb）。

② 人类基因组含有大量的重复DNA（约占50%，见下文），造成了片段组装的困难，因此人类基因组测序联盟没有采用塞莱拉公司的全基因组鸟枪法，而是对染色体大片段克隆（一般100～200kb）进行测序。

③ 国际测序联盟可以将测序任务根据染色体分配到各个测序中心。表17-6列出了这些测序中心。

2000年10月的人类基因组草图版本测序和组装了超过29000个的BAC克隆，总长度达到4.26 Gb（42.6亿碱基对），原始鸟枪测序数据就达到了23Gb。

在人类基因组计划的早期，人们以为DNA测序技术已经能够胜任这样的巨大工程，然而随着计划的进行，DNA测序的Sanger法（第12章）也被不断改进，主要的进步包括（Green的综述，2001）：

① DNA分子自动检测技术的发展，例如出现了基于毛细管电泳的自动测序仪；

② 热稳定性聚合酶的改进；

③ 双脱氧核苷酸的末端荧光标记。

最重要的进展应该是20世纪80年代末自动测序仪的发明。

NCBI的重叠群组装和注释过程见http://www.ncbi.nlm.nih.gov/genome/guide/build.html。

国际测序联盟获得基因组草图的过程是：先挑选BAC克隆，然后测出其原始序列，再对各个克隆的序列进行组装。挑选克隆也是鸟枪测序的一部分，大部分基因文库都包含BAC克隆或P1人工克隆（PAC），它们的DNA来自于匿名捐赠者。

通过链接BAC等大片段克隆的测序结果，序列数据被组装成整合的序列草图（表17-7）。图17-14是该过程的一个例子。

表 17-6 加入人类基因组测序联盟的20个研究机构

测序中心	地点	网址
Baylor College of Medicine	Houston, Texas	http://www.hgsc.bcm.tmc.edu/
Beijing Human Genome Center, Institute of Genetics, Chinese Academy of Sciences	Beijing, China	http://hgc.igtp.ac.cn/
Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center	Cold Spring Harbor, New York	http://nucleus.cshl.org/genseq/lita_annenberg_hazen_genome_cent.htm
Gesellschaft fur Biotechnologische Forschung mbH	Braunschweig, Germany	http://genome.gbf.de/
Genoscope	Evry, France	http://www.genoscope.cns.fr/externe/English/Projets/projets.html
Genome Therapeutics Corporation	Waltham, MA	http://www.genomecorp.com/
Institute for Molecular Biotechnology	Jena, Germany	http://genome.imb-jena.de/
Joint Genome Institute, U.S. Department of Energy	Walnut Creek, California	http://www.jgi.doe.gov/
Keio University	Tokyo, Japan	http://www-alis.tokyo.jst.go.jp/HGS/top.pl
Max Planck Institute for Molecular Genetics	Berlin, Germany	http://seq.mpimg-berlin-dahlem.mpg.de/
Multimegabase Sequencing Center, Institute for Systems Biology	Seattle, Washington	http://www.systemsbiology.org/
RIKEN Genomic Sciences Center	Saitama, Japan	http://hgp.gsc.riken.go.jp/
The Sanger Centre	Hinxton, United Kingdom	http://www.sanger.ac.uk/HGP/
Stanford Genome Technology Center	Palo Alto, California	http://www-sequence.stanford.edu/
Stanford Human Genome Center	Palo Alto, California	http://shgc.stanford.edu/
University of Oklahoma's Advanced Center for Genome Technology	Norman, Oklahoma	http://www.genome.ou.edu/
University of Texas Southwestern Medical Center	Dallas, Texas	http://www3.utsouthwestern.edu/index.htm
University of washington Genome Center	Seattle Washingron	http://www.genome.washington.edu/UWGC/
Washington University Genome Sequencing Center	St. Louis, Missouri	http://genome.wustl.edu/gsc/
Whitehead Institute for Biomedical Research, MIT	Cambridge, Massachusetts	http://www-genome.wi.mit.edu/

表 17-7 NCBI Build 30的不同长度重叠群的数量（2002年10月）

长度范围/kb	数量	总长/kb	含量/%
<30	30	469	0.01
30～100	113	7088	0.24
100～250	547	92587	3.21
250～500	218	78223	2.71
500～1000	227	167372	5.81
1000～5000	479	1079538	37.52
>5000	122	1451804	50.46

注：见 http://www.ncbi.nlm.nih.gov/genome/guide/human/Hs Stats.html。

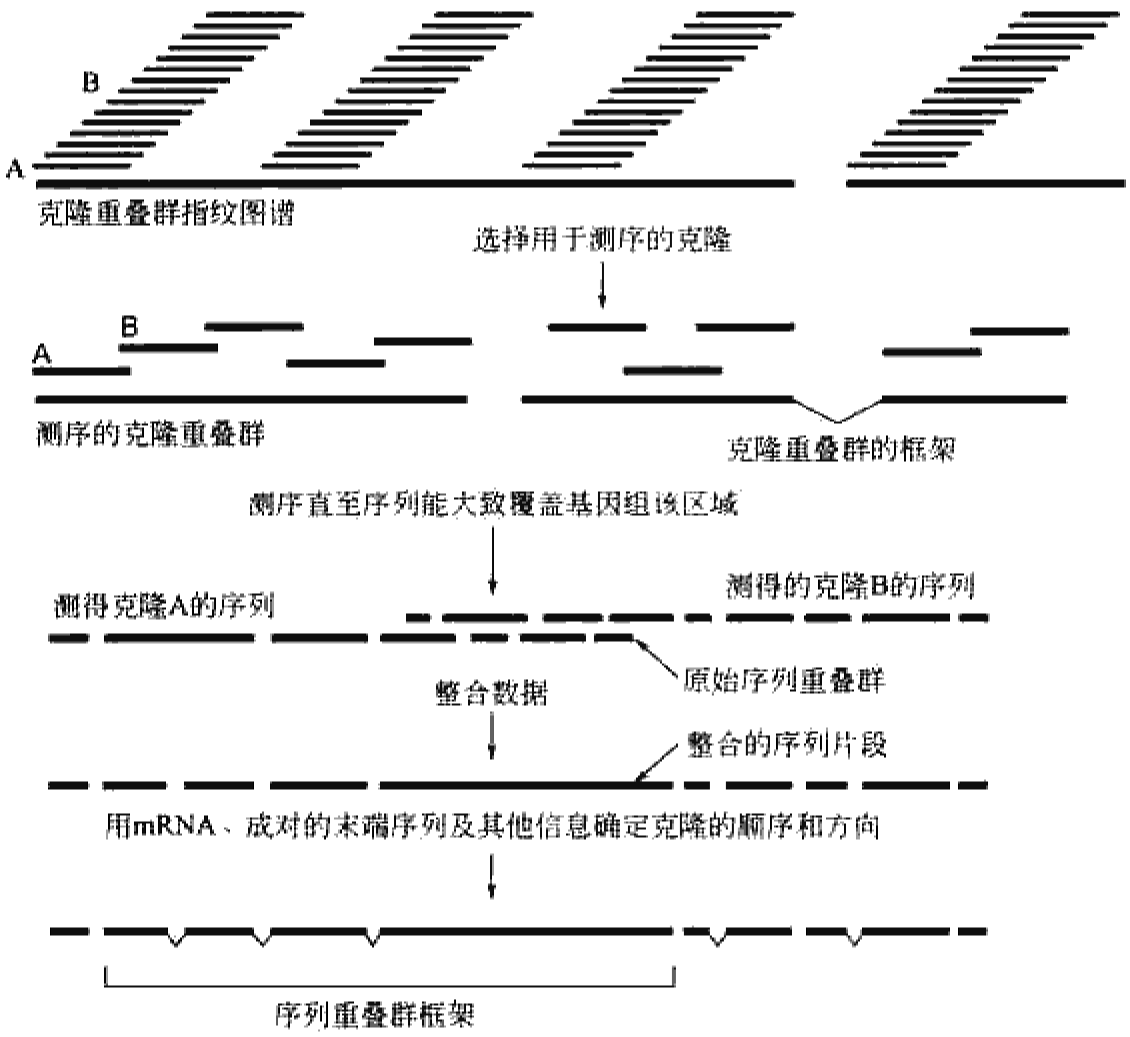


图 17-14 人类基因组的克隆和序列组装。建立基于限制性酶切模式的克隆指纹图谱用以选择相关区域重叠的克隆。这些重叠群被组装成序列框架，确定每一个克隆在基因组中的顺序和方向（IHGSC，2001）。授权使用

17.4.3 测序的一些特点

基因组测序草图包含已完成的数据、草稿和未处理的数据。这些数据中序列片段的连续程度是评价测序的一个重要指标。一个克隆或重叠群的平均长度并不足以衡量基因组被测序和组装的程度，因此采用 N50 长度来表示。N50 长度是指覆盖 50%所有核苷酸的最大序列重叠群长度。在人类基因组草图测序过程中，一半的核苷酸序列存在于一个至少 8.4Mb 的指纹克隆重叠群中（表 17-8）。

表 17-8 基因组序列草图的连续性

克　隆	长度(L)	克　隆	长度(L)
起始测序重叠群	21.7kb	序列重叠群	82kb
序列-重叠群框架	274kb	已测序的克隆重叠群	826kb
指纹克隆重叠群	8.4Mb		

注：1. 用 N50 统计方法得到的结果，即覆盖 50%核苷酸的不同克隆长度。
2. 引自 IHGSC（2001）。

基因测序的质量由缺口的数目和核苷酸的精度来衡量。约 91%未完成的草图序列错误率在 1/10000 以下（PHRAP 分数>40）。

> 关于 N50 统计方法的概叠可在 UCSC 基因组浏览器上找到：http://genome.ucsc.edu/goldenPath/stats.html。

基因测序中另一种基本的错误是组装的错误，问题往往出在多拷贝的重复 DNA 片段或是基因组的重复区域（见下文；Eichler，2001）。比较 IHGSC 和塞莱拉的基因组序列就会发现由于不同的组装而造成的差别（Li 等，2002）。

人类基因组染色体序列的大小见表 17-9，表中常染色体按大小排列。其中最先完成测序的染色体是第 22 号染色体（Dunham 等，1999）、第 21 号染色体

(Hattori 等，2000) 和第 20 号染色体 (Deloukas 等，2001)，它们都是较小的染色体。

表 17-9 人类基因组染色体序列的大小

染色体	大小/Mb	重叠群数目	完成比例/%
1	263	397	32
2	255	436	55
3	214	147	20
4	203	247	21
5	194	431	40
6	183	131	86
7	171	265	79
8	155	85	13
9	145	206	47
10	144	248	47
11	144	183	43
12	143	283	53
13	98	107	84
14	93	69	86
15	89	64	26
16	98	156	34
17	92	138	38
18	85	59	15
19	67	185	61
20	72	9	82
21	34	5	99.5
22	34	12	97
X	164	257	71
Y	35	6	64
all	3175		

注：1. 仅指染色体常染色质的大小。
2. 引自 IHGSC (2001) 及 NCBI (2001 年 10 月)。

17.4.4 人类基因组概貌

通过了解人类基因组的核酸序列，我们就可以概括它的一些主要特点，包括：

PHRAP 分数 40 表示的错误率是 $10^{-40/10}$，或 99.99%的精确度 (见第 12 章)。

① GC 含量分布；
② CpG 岛和重组率；
③ 重复序列含量；
④ 基因含量。

下面我们将逐一介绍这四个特征。使用 NCBI、Ensembl 和 UCSC 的基因组浏览器可以对基因组概貌进行从单个核苷酸到整个染色体各个水平的研究。

17.4.4.1 GC 含量的长程变化

人类基因组的平均 GC 含量是 41%，但是存在一些相对富含 GC 和缺乏 GC 的区域。总体 GC 含量直方图 (20kb 窗口) 显示向右连续下降的总体趋势 (图 17-15)。58%的窗口 GC 含量低于平均值，42%高于平均值，包含一个高 GC 区域的长尾区。

Giorgio Bernardi 和他的同事们提出哺乳动物基因组是由一些大的 DNA 片段 (如>300kb) 镶嵌而成。这些片段被称为"等容线"，它们组分均一，可以分为低 GC 家族 (L1、L2) 和高 GC 家族 (H1、H2、H3)。然而 IHGSC (2001) 报告未鉴别出严格意义上的等容线，Haring

和 Kypr（2001）在人类 21 号、22 号染色体上也未发现等容线。但 Bernardi 和他的同事们随后的分析（Bernardi，2001；Pavlicek，2002）仍支持人类基因组是由不同 GC 含量的片段镶嵌组成。造成这一分歧的部分原因是分析时选定的基因组 DNA 窗口的大小不同。

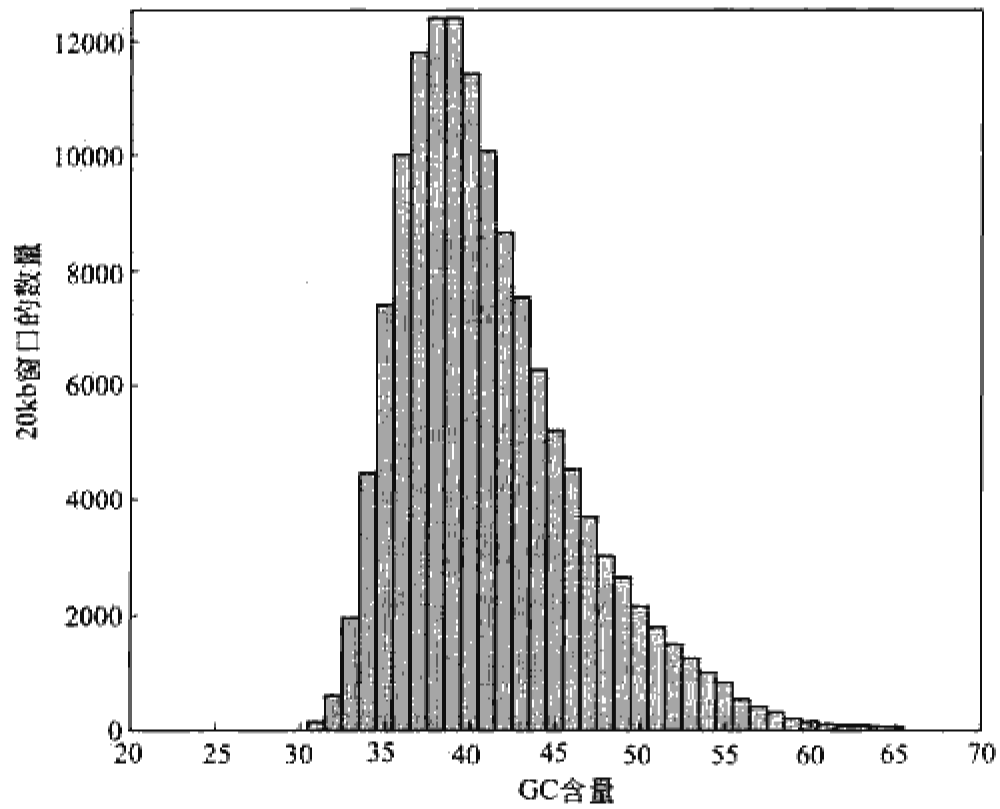


图 17-15 基因组草图序列 GC 含量直方图（20kb 窗口）。注意分布向右连续下降，平均 GC 含量为 41%。摘自 IGHSC（2001），已获授权

17.4.4.2 CpG 岛

二核苷酸的 CpG 在基因组 DNA 中的含量非常低，出现率仅为预计值的 1/5。因为大部分 CpG 的胞嘧啶被甲基化，然后脱去氨基。然而在基因组中仍含有许多“CpG 岛”，它们常常与看家基因的启动子和外显子区相联系（Gardiner-Garden 和 Frommer，1987）。这些 CpG 岛参与基因沉默、基因组印迹（Tycko 和 Morison，2002）以及 X 染色体失活（Avner 和 Heard，2001）等过程。

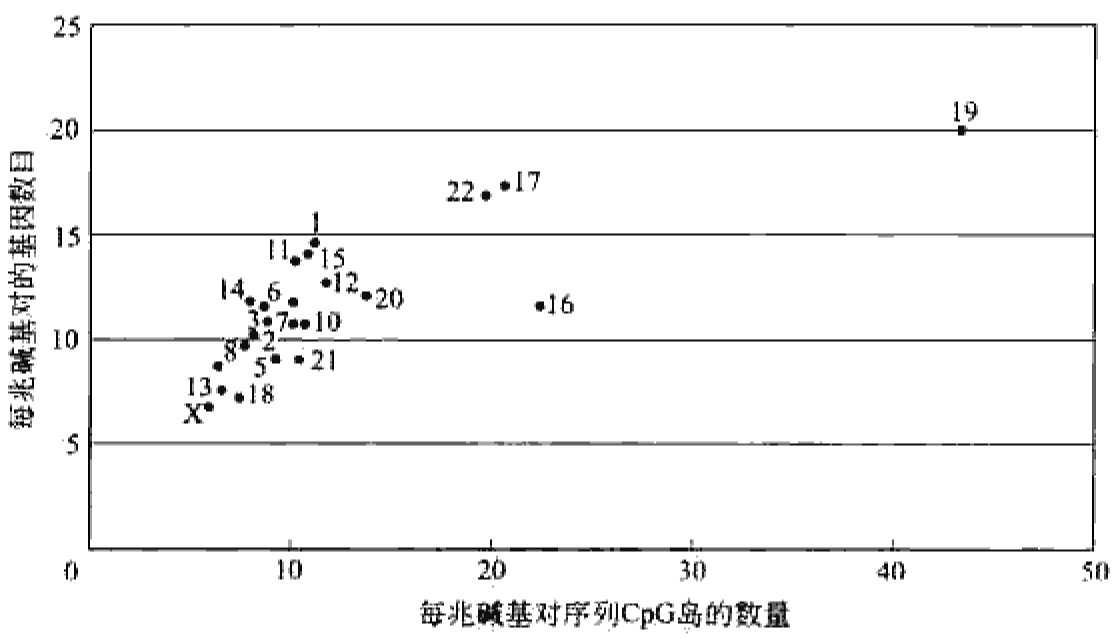


图 17-16 每个染色体每兆碱基对序列中 CpG 岛的数目和针对每兆碱基对序列的基因数目。注意基因最丰富的 19 号染色体每兆碱基对有最多的 CpG 岛数目。摘自 IHGSC（2001），已获授权

在 NCBI、Ensembl 和 UCSC 基因组浏览器网站上都可以显示预测的 CpG 岛（例如表 17-2 和图 17-2）。根据 IHGSC 的报告（2001），人类基因组中有 50267 个预测的 CpG 岛，用 RepeatMasker 去除重复 DNA 序列后还有 28890 个 CpG 岛（数量的下降反映了 *Alu* 重复元件的高 GC 值，见图 16-8）。大部分染色体上每兆碱基对的 DNA 序列就有 5～15 个 CpG 岛，在 19 号染色体（基因密度最大的染色体）上每兆碱基对有 43 个 CpG 岛（图 17-16）。

可以在 NCBI、Esembl 或 UCSC 基因组浏览器中查看各个染色体上的 GC 含量。例如，在 Ensembl 浏览器中（图 17-7）点击"decorations"可增加显示 GC 含量的图层。

等容线的 L（轻）、H（重）划分涉及基因组 DNA 片段在氯化铯梯度离心时的沉淀行为。经过梯度离心，不同 GC 百分含量的 DNA 片段会被分到不同的层中。

17.4.4.3 遗传距离与物理距离的比较

可以通过比较染色体的遗传图和物理图来估计每个核苷酸的重组率（Yu 等，2001）。遗传图即连锁图，是基于减数分裂重组的染色体图。在减数分裂中细胞两份染色体拷贝减为一份，同源重组（交换 DNA）过程发生。遗传图用基因的重组概率来描述 DNA 序列的相对距离，以 cM（centimorgans）为单位。1cM 相当于 1%的重组率。

与遗传图不同，物理图描述的是核苷酸序列在染色体上的实际物理位置。人类基因组序列草图的完成使遗传图和物理图的比较成为可能。

图 17-17 显示了人类第 12 号染色体序列的遗传距离（y 轴，单位 cM）和物理距离（x 轴，单位 Mb）（IGHSC，2001）。有两个主要的结论：第一，染色体着丝粒部分重组受到限制，而靠近端粒部分重组率增加巨大，这种现象在男性中尤其明显。第二，长染色体臂倾向于 1cM/Mb 的平均重组率，而短臂具有高得多的重组率（约 2cM/Mb）。整个基因组的重组率分布为 0～9cM/Mb（Yu 等，2001）。研究还发现了 19 个重组"沙漠"（长度超过 5Mb，平均重组率低于 0.3cM/Mb）和 12 个重组"森林"（长度超过 6Mb，平均重组率高于 3.0cM/Mb）。

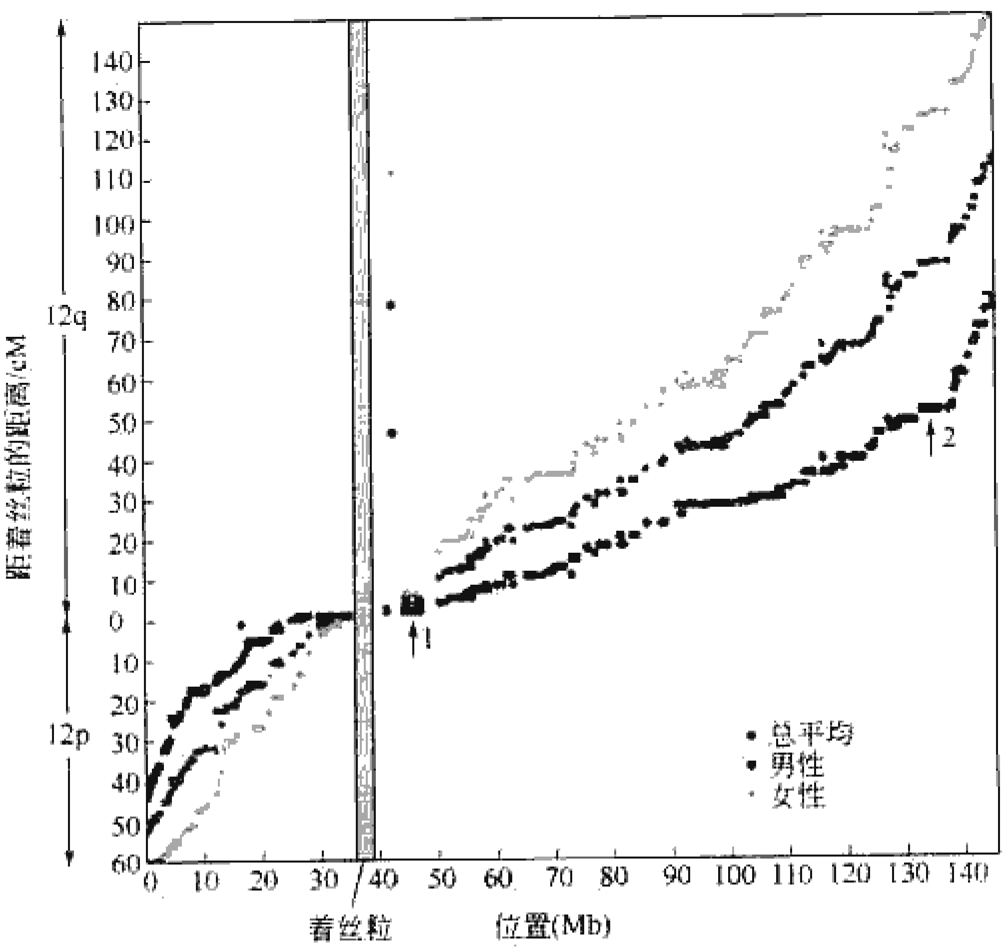


图 17-17 人类第 12 号染色体序列遗传距离（y 轴，单位 cM）与物理距离（x 轴，单位 Mb）的比较。注意着丝粒部分重组率较低（箭头 1），而靠近端粒部分（各染色体的末端部分）重组率高，尤其在男性减数分裂图上（箭头 2）摘自 IHGSC（2001），授权使用

17.4.5 人类基因组重复序列的含量

重复 DNA 序列约占据人类基因组的 50%以上，它们的来源和功能仍是引人注目的问题。到底有几种不同的重复序列？它们在何时何处产生？它们为什么混杂在我们的基因组中

不断生长？它们的复制有意义吗？人类基因组计划的成果之一就是让我们开始了解基因组重复序列存在的区域及其性质。

> 基因沉默与转录抑制有关。例如MeCP2是一个与甲基化的CpG岛结合的蛋白（图8-3），能结合其他蛋白（如组蛋白脱乙酰酶），引起染色体结构的改变从而抑制转录。X染色体上表达MeCP2的*MeCP2*基因的突变会引起Rettsydrome（Amir等，1999），该疾病会引起女孩特殊的神经综合征，包括手自主运动、抓取功能丧失和类似孤独症的行为（第18章）。

在人类基因组中有五类重复DNA序列（IHGSC，2001；Jurka，1998）。

① 散在插入重复（转座子引起的重复）。

② 过程假基因：失活或部分失活编码基因的反转录拷贝。

③ 简单序列重复：微卫星和小卫星，包括类似$(A)_n$、$(CA)_n$、$(CGG)_n$的短序列。

④ 片段性重复，包括大约10～300kb的片段，这些片段从基因组的一个区域拷贝到另一区域。

⑤ 串联重复序列区，例如着丝粒、端粒和核糖体的基因簇。

我们将逐一介绍各类重复序列。

17.4.5.1 转座子来源的重复序列

45%的人类基因组由转座子来源的重复序列构成，这很令人惊奇。它们也被叫做散在重复序列。许多转座子来源的重复序列在久远的过去（数亿年前）就在人类基因组上复制增殖，它们在序列上变化太大而不能被辨认出来，所以45%很可能只是一个较低的估计。转座子来源的重复序列可分为四类（Jurka，1998；Ostertag和Kazazian，2001）：

> 基因组印迹是指母系和父系等位基因的差异表达情况。Tycko和Morison（2002）提供了一个印迹基因数据库：http://www.otago.ac.nz/IGC。

① 占人类基因组21%的长散在元件LINE；

② 占人类基因组13%的短散在元件SINE；

③ LTR转座子，占人类基因组8%；

④ DNA转座子，约占人类基因组3%。

图17-18显示了这些重复序列的结构和它们在人类基因组中的拷贝数。LINE、SINE和LTR转座子都是具有编码反转录酶活性的反转录转座子，它们通过产生RNA中间体整合入基因组。与它们不同，DNA转座子具有反向末端重复序列，并编码类似细菌转座子的转座酶。

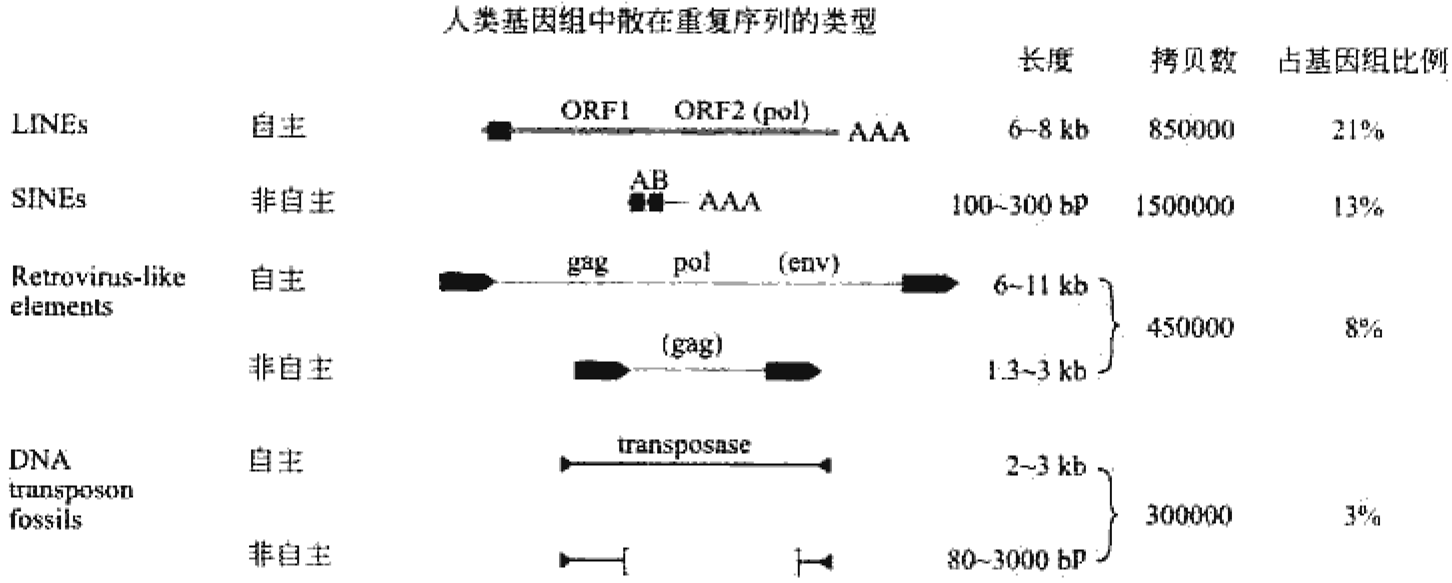


图17-18 人类基因组中有四种可转座元件：LINE、SINE、LTR转座子和DNA转座子。引自IHGSC（2001），授权使用

反转录转座子可以更细地分为自主的反转录转座子（含有它们转座所需的酶的编码序列）和非自主反转录转座子（依靠外源的酶，如宿主细胞的DNA修复酶）。最常见的非自主反转录转座子是*Alu*元件。

X染色体失活是指女性个体的细胞选择性失活来源于父方或母方的一条X染色体的现象，称为剂量补偿效应（Avner和Heard，2001）。

散在重复序列占人类基因组的比例比在其他真核生物基因组中大得多（表17-10），拷贝总数估计有300万，它们的存在为对分子进化的研究提供了有力的工具。每一个重复元件，即使已经失活，都是一个可用来研究物种内与物种间基因组演化的"化石记录"；转座子随机独立地积累突变，因此我们可以通过转座子间的多序列比对来计算序列差异度，并将转座子的进化作为分子钟，分子钟的校准可基于物种分化的时间，如人类和古猿分化是在2300万年前。根据这种系统发生学分析已得到一些结论（IHGSC，2001）。

表17-10 四种真核基因组的散在重复序列

项目	人		果蝇		线虫		芥菜	
	碱基百分比	家族近似值	碱基百分比	家族近似值	碱基百分比	家族近似值	碱基百分比	家族近似值
LINE/SINE	33.4%	6	0.7%	20	0.4%	10	0.5%	10
LTR	8.1%	100	1.5%	50	0%	4	4.8%	70
DNA	2.8%	60	0.7%	20	5.3%	80	5.1%	80
总计	44.4%	170	3.1%	90	6.5%	90	10.5%	160

注：引自IHGSC（2001），授权使用。

NCBI、Ensembl和UCSC基因组浏览器都可以显示遗传图和物理图。例如，NCBI在它的图显示器中提供遗传图的3种显示方式（表17-2，图17-2）。

① 人类基因组中大部分散在重复序列是很古老的，在哺乳亚纲出现的年代（1亿年前）之前就已产生。这些元件仅是缓慢地从基因组中移除。

② SINE和LINE在很久以前（至少1.5亿年前）就已经存在。

③ 在过去的5000万年中，没有证据显示人类基因组中有DNA转座子活动，因此，它们已经成为化石。

④ 相对于其他真核生物，人类基因组具有更大比例的散在重复序列和许多更古老的重复序列（图17-19）。

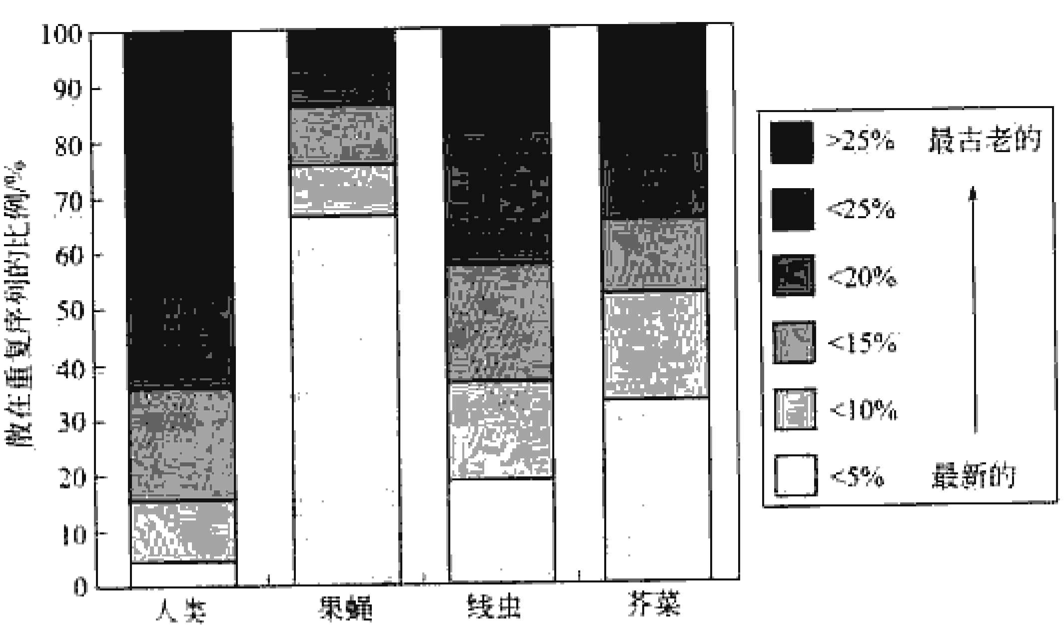


图17-19 真核生物基因组中散在重复序列的年龄比较。年代较近的重复序列在人类中只有很小的比例。引自IHGSC（2001），授权使用

17.4.5.2 简单序列重复

简单重复序列是完美的或略不完美的特殊*k*-mer（*k*个碱基）的串联重复。具有短的重复单位（1～12bp）的简单序列重复称为微卫星，具有较长的重复单位的序列（13～500bp）

通常称为小卫星（Toth 等，2000）。

微卫星和小卫星占人类基因组的 3%（IHGSC，2001）。最普遍的重复长度见表 17-11。最常见的重复单位是二核苷酸 AC、AT 和 AG，例子在 RepeatMasker 程序介绍中可见（第 16 章）。

表 17-11 人类基因组中的简单重复序列（微卫星）

重复长度	每兆碱基对平均碱基数	每兆碱基对中简单重复序列(SSR)的平均数目
1	1660	33.7
2	5046	43.1
3	1013	11.8
4	3383	32.5
5	2686	17.6
6	1376	15.2
7	906	8.4
8	1139	11.1
9	900	8.6
10	1576	8.6
11	770	8.7

注：来源于 IHGSC（2001）。SSR 表示简单序列重复。

17.4.5.3 片段性重复

片段性重复与人类基因组的 1～200kb 序列被复制到基因组上一个或多个位置有关。已完成的人类基因组序列中约 3.6% 由片段性重复组成，一般长 10～50kb（Bailey 等，2001）。许多这种重复是在相对较近的时间发生的，因为内含子和编码区都有高的保守性（对于古老的复制事件，重复片段的内含子区域保守性会比较低）。片段性重复可以在同一条染色体内，也可以在染色体间发生。着丝粒包含了大量的染色体间片段性重复，例如 22 号染色体长达 1.5Mb 毗邻着丝粒区域的 90%序列由染色体间复制序列组成（图 17-20）。较小范围的片段性重复也发生在接近端粒的区域。

估计散在重复的数目可使用 RepeatMasker 搜索 RepBase 用软件 RepeatMasker 搜索 RepBase 数据库可以估计散在重复序列的数目（见第 16 章）。

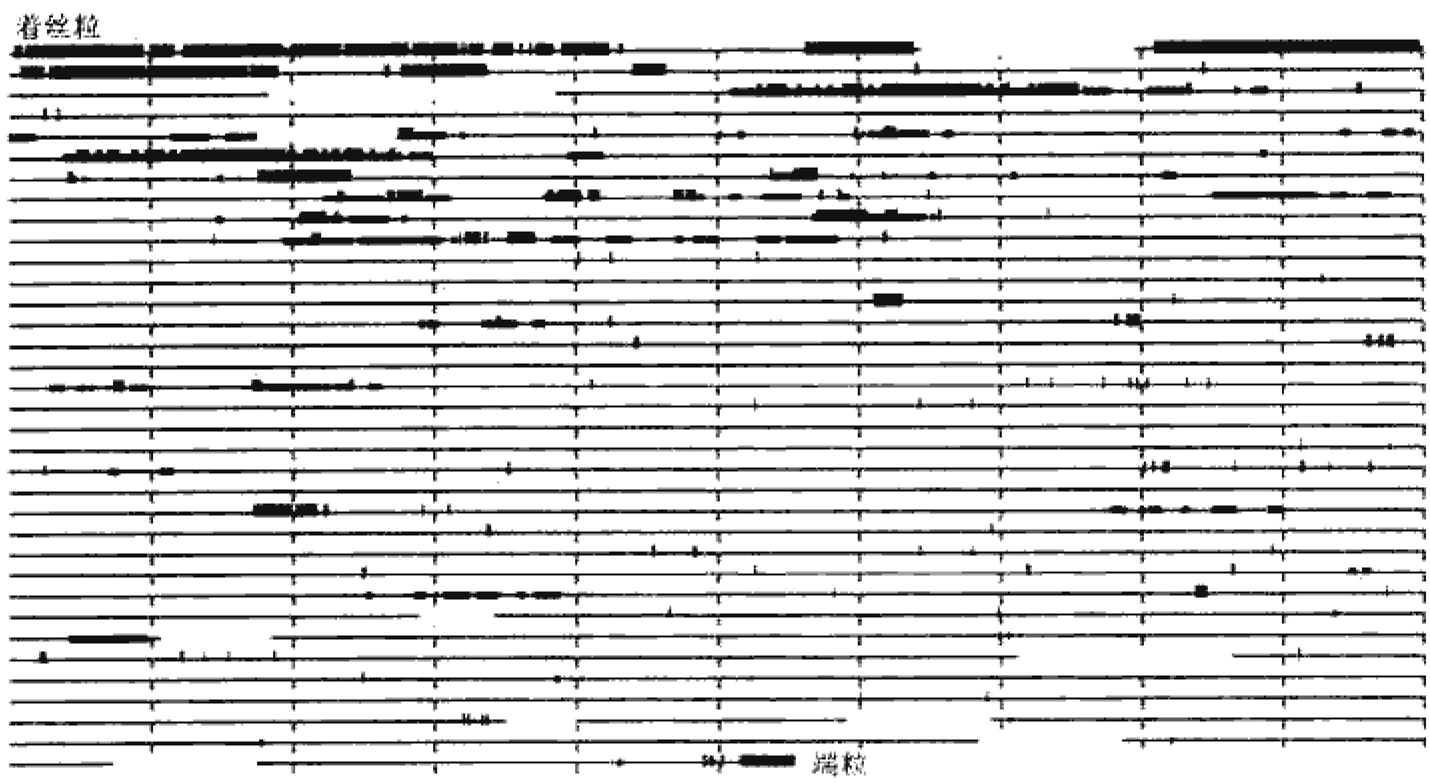


图 17-20 22 号染色体的着丝粒包含大量的染色体间片段性重复。图上显示染色体内（黑色）和染色体间（红色）重复片段的大小和位置。每条水平线代表 1Mb（每个标记 100kb 间隔）。着丝粒在左上，端粒在右下。摘自 IHGSC（2001），授权使用

17.4.6 人类基因组基因含量

由于基因在生物学中的重要地位，研究人类基因组的基因含量成为人们关注的焦点。然而，辨别基因是在人类基因组序列识别中最难的部分（第16章），是项具有挑战性的工作，原因如下。

之所以称为 *Alu* 元件是因为内切酶 Alu Ⅰ 能把该元件的序列从中间切开。在小鼠中同样的元件称为 B1。

① 外显子的平均长度只有50个密码子（150个核苷酸），这样小的序列难以明确地识别。

② 外显子被内含子隔开，有些间隔达数万碱基对。极端的例子如人类 *dystrophin* 基因全长超过2.4Mb，这相当于典型原核生物整个基因组的长度。因此，获取cDNA仍是识别基因的重要手段。

③ 存在许多假基因，它们和编码功能蛋白的真基因难以区分。

④ 我们对非编码基因还很不了解（见第16章和下文）。

17.4.6.1 非编码 RNA

人类基因中有许多并不编码蛋白质，而以非编码RNA为其产物。非编码RNA在基因组DNA中难以识别，因为它们较短小、缺乏开放阅读框，也没有poly（A），基因寻找算法难以检测，也不存在于cDNA文库中。这类非编码RNA包括：

Annemarie Poustka 和他的同事们最近对500个包含完整阅读框的新的人类cDNA测序。他们测了超过30000个cDNA克隆的5′末端（Wiemann 等，2001）。结果数据在 http://www.rzpd.de/。

① 转运RNA（tRNA），用于将mRNA翻译成蛋白质的氨基酸序列；

② 核糖体RNA（rRNA），用于mRNA的翻译；

③ 小核仁RNA（snoRNA），是核仁中RNA加工过程所必需的；

④ 小核RNA（snRNA），剪接体行使功能所必需的。

在人类基因组草图中数百个非编码RNA已被识别（表17-12）。其中tRNA基因最多（有497个），并有324个tRNA来源的假基因。各tRNA基因所对应的人类遗传密码已可阐明。新版本的遗传密码图包含了各个氨基酸密码子的使用频率和对应tRNA基因的数目。此外，得到的人类tRNA基因的总量还可以和其他真核生物的作比较（表17-13）。

表 17-12 人类基因组中的非编码基因

RNA 基因	非编码基因数目	相关基因数目	功　能
tRNA	497	324	蛋白合成
SSU(18S)RNA	0	40	蛋白合成
5.8S rRNA	1	11	蛋白合成
LSU(28S)rRNA	0	181	蛋白合成
5S RNA	4	520	蛋白合成
U1	16	134	剪接体成分
U2	6	94	剪接体成分
U4	4	87	剪接体成分
U4 atac	1	20	Minor(*U11*/*U12*)剪接体成分
U5	1	31	剪接体成分
U6	44	1135	剪接体成分
U6 atac	4	32	Minor(*U11*/*U12*)剪接体组分

续表

RNA 基因	非编码基因数目	相关基因数目	功 能
U7	1	3	组蛋白 mRNA3′加工
U11	0	6	Minor(U11/U12)剪接体组分
U12	1	0	Minor(U11/U12)剪接体组分
SRP(7SL)RNA	3	773	信号识别分子的组分
RNAse P	1	2	tRNA 5′端加工
RNAse MRP	1	6	rRNA 加工
Telomerase RNA	1	4	端粒增长的模板
hY1	1	353	Ro RNP 的组分，功能未知
hY3	25	414	Ro RNP 的组分，功能未知
hY4	3	115	Ro RNP 的组分，功能未知
hY5(4.5S RNA)	1	9	Ro RNP 的组分，功能未知
Vault RNAs	3	1	13Mda Vault RNP 的组分
7SK	1	330	未知
H19	1	2	未知
Xist	1	0	X 染色体失活的起始
Known C/D snoRNAs	69	558	rRNA 前体的加工或 rRNA 位点特异性甲基化
Known H/ACA snoRNAs	15	87	rRNA 前体的加工或 rRNA 位点特异性假尿苷酸化

注：引自 IHGSC (2001)，授权使用。

表 17-13 多种生物的 tRNA 基因估计数目

生 物	标准的 tRNA	SeCys tRNA
人	497	1
线虫	584	1
果蝇	284	1
酵母菌	273	0
詹氏甲烷球菌	36	1
大肠杆菌	86	1

注：SeCys 表示硒半胱氨酸。

17.4.6.2 蛋白编码基因

蛋白编码基因具有外显子、内含子和调控元件，表 17-14 总结了这些基本特征。人类基因平均编码序列长度为 1340bp (IHGSC，2001)，线虫的是 1311bp，而果蝇为 1497bp。线虫和果蝇较长外显子的比例比人类的大一些［图 17-21(a) 曲线的尾部］，但这三个物种的中间外显子大约都是 50～200bp［图 17-21(a)］，而且人类内含子长度的变化要大得多［图 17-21(b)，图 17-21(c)］，因而人类基因长度比线虫和果蝇有更大的变化范围。

表 17-14 人类基因的特性

特 征	大小(中位数)	大小(平均值)
中间外显子	122 bp	145 bp
外显子数目	7	8.8
内含子	1023 bp	3365 bp
3′-UTR	400 bp	770 bp
5′-UTR	240 bp	300 bp
编码序列	1100 bp	1340 bp
编码肽链长度	367 aa	447 aa
基因总长	14 kb	27 kb

注：1. aa 表示氨基酸。
2. 引自 IHGSC (2001)。

蛋白编码基因与高的GC含量相关联（图17-22）。人类基因组总的GC含量约是41%，其

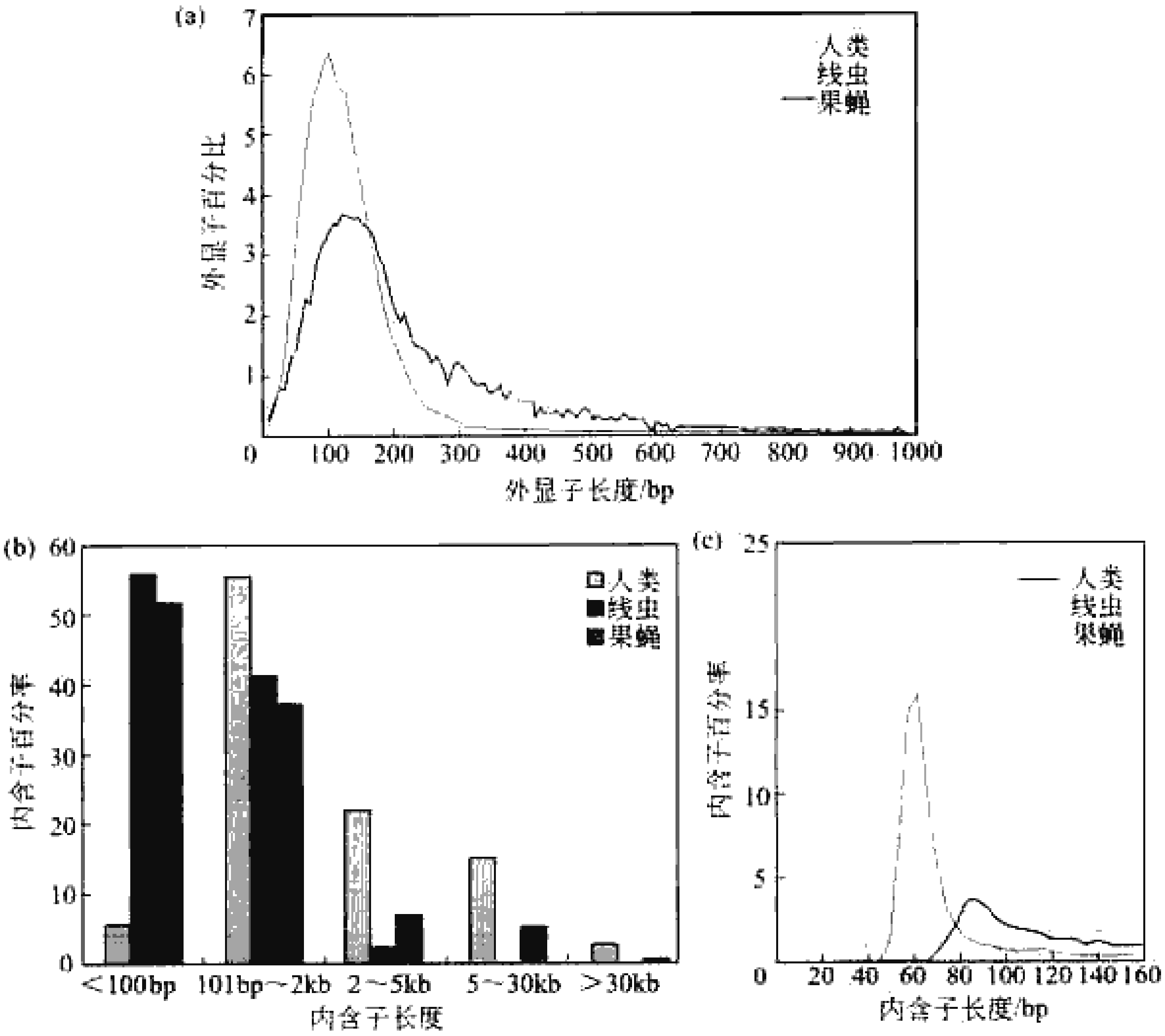


图17-21 在人类、线虫及果蝇基因组中外显子和内含子的分布。(a) 外显子，(b) 内含子，(c) 短内含子［从(b)放大］。引自IHGSC（2001），授权使用

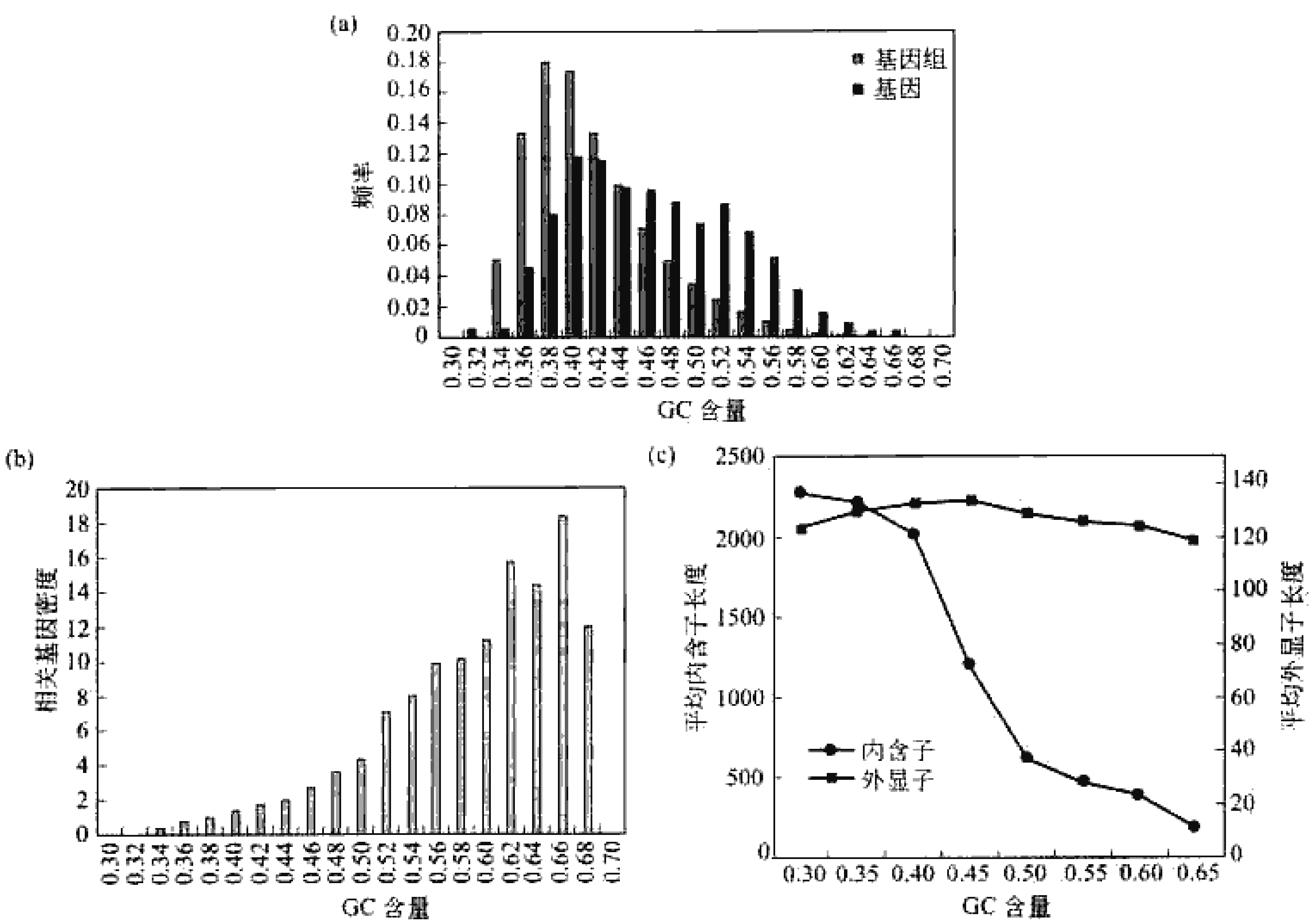


图17-22 (a) 基因组和基因中GC含量的分布说明蛋白编码基因和高GC含量相关联。(b) 基因密度作为GC含量的函数［基因密度通过（a）中数据的比例得到］，随GC含量的上升，其迅速增加。(c) 平均外显子长度不受GC含量的影响，但内含子随GC含量上升而迅速变短。引自IHGSC（2001），授权使用

中已知基因（有 RefSeq ID）的 GC 含量较高［图 17-22(a)］。当 GC 含量从 30%增加到 50%时基因密度增加了 10 倍［图 17-22(b)］。

最长的编码序列是 *titin*（80780bp；编号 NM_003319)。titin 基因在染色体 2q24.3，有 178 个外显子，编码一个含 26926 个氨基酸的肌蛋白（约 3000ku)。与之相反的是一个典型的由 1340bp mRNA 翻译的蛋白，约是 50ku。

为了编写所有蛋白编码基因和它们的蛋白产物的目录，IHGSC（2001）创建了一个完整基因索引（IGI）和相应的完整蛋白索引（IPI)。根据欧洲生物信息学研究中心（EBI）的蛋白组分析结果，人类蛋白组含 33817 个蛋白。Ensembl 预测显示了 15 个最普遍的蛋白家族（表 17-15)、最普遍的结构域（表 17-16）以及最普遍的重复区（表 17-17)。

表 17-15 人类 15 种最常见的蛋白家族

Inter Pro 编号	匹配蛋白数量	名 称
IPR000276	837	类视紫红质 GPCR 超家族
IPR000719	685	蛋白激酶
IPR001909	305	KRAB 盒
IPR001806	191	Ras GTP 酶超家族
IPR005821	153	离子转运蛋白
IPR000387	129	酪氨酸特异性蛋白磷酸化酶
IPR001254	126	丝氨酸蛋白酶，胰蛋白酶家族
IPR000379	107	各种脂酶
IPR001664	83	中间纤维蛋白
IPR001128	79	细胞色素 P450
IPR000910	78	HMG1/2 盒
IPR000832	73	G 蛋白耦联受体家族 2(类促胰液素)
IPR002198	72	短链脱氢酶/还原酶
IPR005828	72	底物运输蛋白
IPR001394	67	泛素硫酯酶，家族 II

注：来源于 http://www.ebi.ac.uk/proteome/，2002 年 11 月。

表 17-16 人类蛋白 15 种最普遍的结构域

InterPro 编号	结构域数量(蛋白数量)	名 称
IPR000822	27858(912)	锌指结构(2H2 类型)
IPR003006	2329(853)	免疫球蛋白/主要组织相容性复合物
IPR002965	1525(364)	富含脯氨酸伸展素
IPR001841	1090(361)	锌指结构 RING
IPR001849	1054(330)	类似 Pleckstrin 区域
IPR000504	1620(317)	RNA 结合区 RNP-1(RNA 识别模体)
IPR006209	2942(299)	类 EGF 结构域
IPR001452	1814(282)	SH3 结构域
IPR002048	1682(263)	钙结合的 EF-hand
IPR001356	1403(251)	同源异型盒
IPR003961	1819(229)	纤黏连蛋白 III 型
IPR001478	1097(208)	PDZ/DHR/GLGF 结构域
IPR000210	498(194)	BTB/POZ 结构域
IPR005225	188(188)	小 GTP 结合蛋白结构域
IPR002126	4168(169)	钙黏蛋白 Cadherin

注：1. 括号中的数字是匹配蛋白的数量。

2. 来源于 http://www.ebi.ac.uk/proteome/，2002 年 11 月。

表 17-17 人类蛋白 15 种最普遍的重复区

InterPro 编号	匹配蛋白数量	名 称
IPR001680	364	G 蛋白 βWD-40 重复
IPR002110	263	锚蛋白
IPR001611	215	富含亮氨酸的重复
IPR001440	173	TPR 重复
IPR000087	97	胶原蛋白三重螺旋结构
IPR001798	95	Kelch 重复
IPR002017	80	Spectrin 重复
IPR000884	76	血小板凝集素Ⅰ型
IPR003659	59	Plexin/semaphorin/integrin
IPR002172	53	低密度脂蛋白受体 A 类,结构域
IPR000225	46	Armadillo 重复
IPR001610	34	PAC 模体
IPR002165	30	丛蛋白 Plexin
IPR000357	28	HEAT 重复
IPR000585	25	Hemopexin 重复

注：来源于 http://www.ebi.ac.uk/proteome/，2002 年 11 月。

17.4.6.3 比较蛋白组分析

比较分析的重要性使之成为基因组学的基本原则。表 17-18 显示了人类的蛋白质与酵母（*S. cerevisiae*）、拟南芥（*A. thaliana*）、线虫（*C. elegans*）和果蝇（*D. melanogaster*）全基因组中蛋白质的比较。IHGSC（2001）根据 InterPro 和 Gene Ontology（GO）的分类分析了这些蛋白的功能类别。图 17-23 显示人类在细胞骨架、转录/翻译、防卫和免疫方面具有更多的蛋白编码基因。

19 号染色体含有最多的 GC 富集区，同时也具有最高的基因密度（26.8 个/Mb）。在基因组中，预测基因的平均密度是 11.1 个/Mb。基因密度最低的是 Y 染色体，只有 6.4 个/Mb 预测基因。

表 17-18 人类与酵母（*S. cerevisiae*）、拟南芥（*A. thaliana*）、线虫（*C. elegans*）和果蝇（*D. melanogaster*）的蛋白质组比较

物 种	蛋白组总蛋白数据	InterPro 中有匹配的蛋白数量	占总蛋白数的比例/%	已有标号的数量	InterPro 中的蛋白记录数
H. sapiens	28469	20530	72.1	5993	3193
M. musculus	21001	15898	75.7	5733	3046
A. thaliana	18064	13684	75.8	4103	2168
C. elegans	20506	14421	70.3	4316	2251
D. melanogaster	13774	10460	75.9	4409	2279
S. cerevisiae	6159	4192	68.1	3578	1816

注：来源于 http://www.ebi.ac.uk/proteome/，2002 年 11 月。

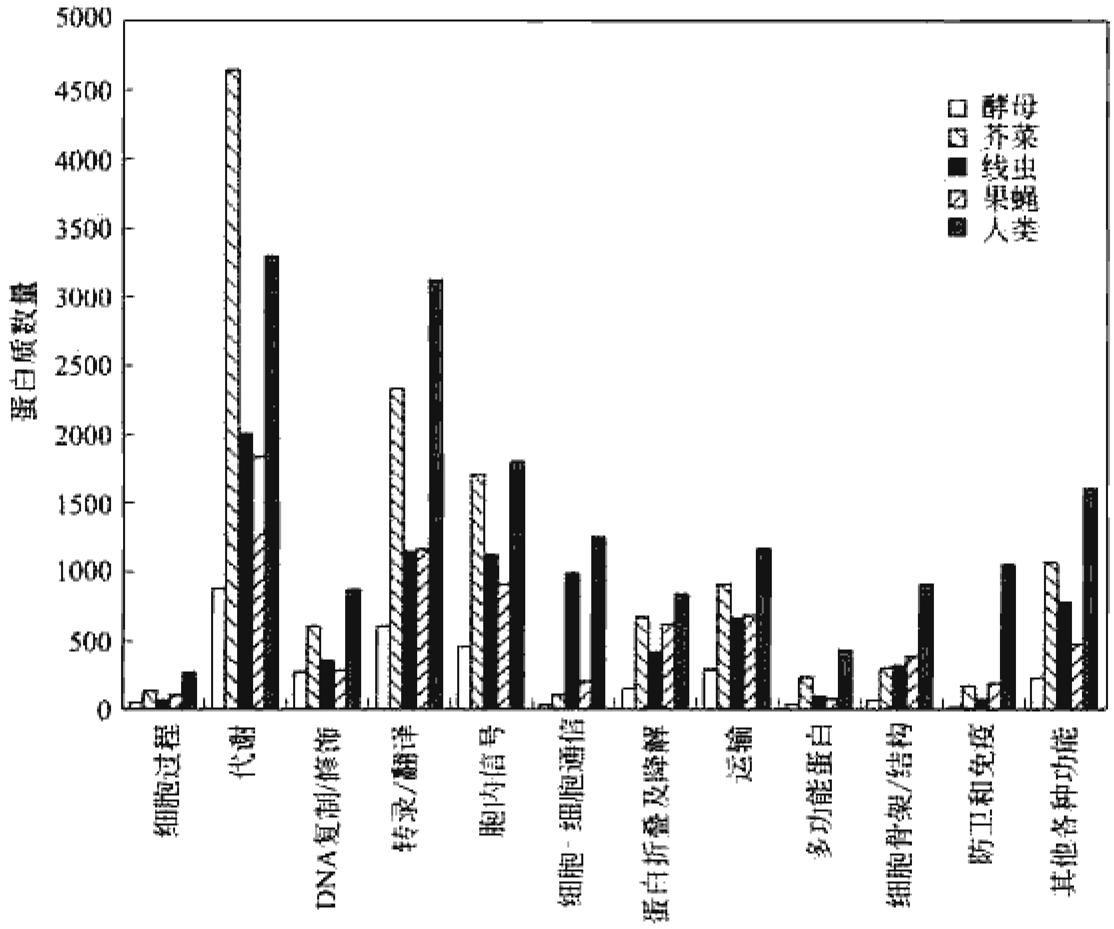


图 17-23　真核生物蛋白的功能类别，有酵母（*S. cerevisiae*）、拟南芥（*A. thaliana*）、线虫（*C. elegans*）、果蝇（*D. melanogaster*）和人类。类别来自于 InterPro（第 10 章）。引自 IHGSC（2001），授权使用

进一步的研究对人类蛋白组中每一个预测的蛋白使用 blastp 在非冗余数据库中搜索，得到如图 17-24 的同源性分布。总体上看 74%的蛋白明显与其他已知蛋白相关。相信随着数据库中序列的增多，人类蛋白和其他真核生物的蛋白的匹配数将会继续增长。

EBI 蛋白质组分析见 http://www.ebi.ac.uk/proteome/。

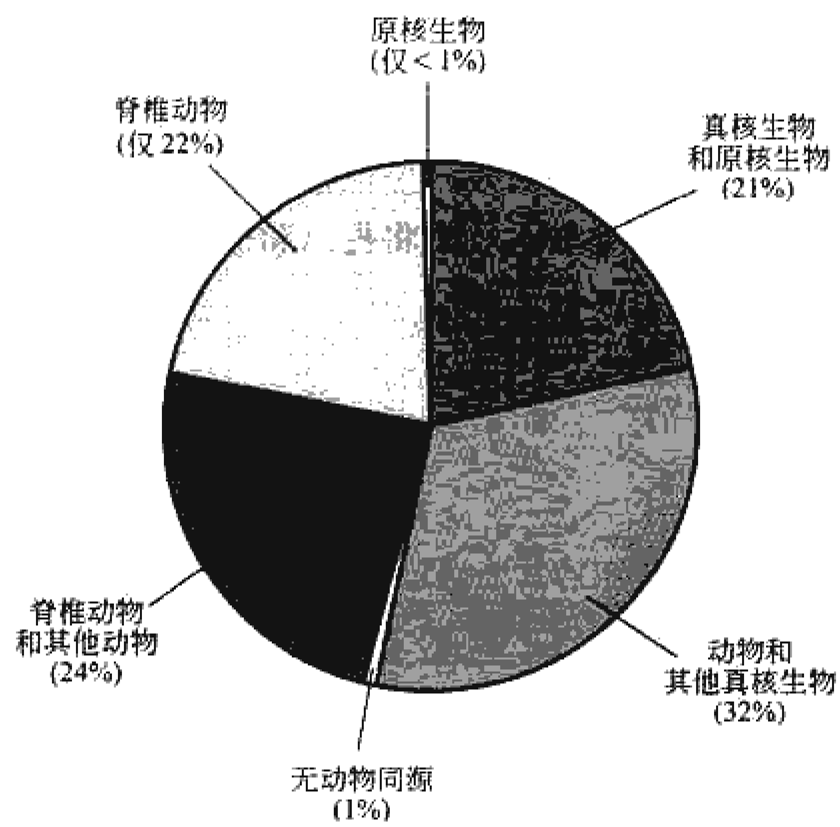


图 17-24　预测的人蛋白质的物种同源性分布。每个蛋白都用 blastp 搜索，*E* 值低于 0.001 的为同源蛋白。用 PSI-BLAST 进行了额外搜索（重复 3 次）。引自 IHGSC（2001），授权使用

17.4.6.4　人类蛋白质组的复杂性

人类蛋白编码基因的数量和其他的多细胞动物、植物相差不大，仅比单细胞真菌大 5 倍，但是人类蛋白组可能要复杂得多，原因如下（IHGSC，2001）。

我们在第 8 章和第 10 章讨论了 GO 协会和 Interpro。

① 相对来说，人类比其他生物有更多的结构域和蛋白家族。

② 人类基因组含有较多旁系同源基因，有产生更多样功能的潜力。

③ 有更多具有复合功能的多结构域蛋白。

④ 人类蛋白组的结构域体系更复杂。

⑤ 人类 RNA 的可变剪切更广泛。

在这些因素协同作用下引起了人类蛋白组的巨大复杂性，也说明了哺乳动物（包括人类）表型复杂性的由来。

17.5 展望

人类基因组测序是国际协作努力数十年的结果，是科学史上的伟大成就。测序的完成基于两个主要的技术进步：①20 世纪 80 年代 DNA 自动测序仪的发明使核苷酸数据可以被大规模地采集；②生物学家和计算机科学家开发了许多计算生物学工具可对序列数据进行分析。在不久的将来，我们可以预见 DNA 测序的进程持续发展，我们将很快能够比较许多个体的全基因组序列，努力揭示基因型和表型的联系。

17.6 常见问题

有许多还未解决的重要问题：

① 我们如何精确地计算蛋白编码基因的数量？

② 如何确定非编码基因的数目？

③ 如何确定基因和蛋白的功能？

④ 我们有着怎样的物种进化历史？

⑤ 个体差异在核苷酸水平上是如何表现的？

因此，当我们看到人类基因组序列的时候，应该把它作为一个开始而不是结束。通过基因组序列，并通过其和其他基因组的比较，我们将提出和面对一系列新的问题。

17.7 网络资源

相关网络资源见表 17-19。

表 17-19 人类基因组序列数据公共资源的其他网址

资源	简述	网址
CBI	查看染色体，图谱和位点	http://ncbi.nlm.nih.gov/genome/guide
橡树岭国家实验室	查看人类基因组	http://compbio.ornl.gov/channel/index.html
RIKEN 东京大学	人类基因组概览	http://hgrep.ims.u-tokyo.ac.jp/
华盛顿大学	人类基因组的克隆和图谱链接	http://genome.wustl.edu/gsc/human/Mapping/

17.8 问题讨论

1. 如果你有测量 5 个个体的全基因组序列的资源和设备，你将如何选择这些个体？ 为什么？

2. 用BLAST搜索可发现酵母（*Saccharomyces cerevisiae*）基因组约在1亿年前发生过复制（duplication）（第15章）。为什么不能直接用这种方法检测人类基因组的大片段复制？是因为在人类基因组中这种现象没有发生过或是不明显？还是因为我们缺乏分析这种大规模基因组变化的工具？为了更深入地讨论人类基因组的大片段复制问题，可以参考Evan Eichler（2001）的一篇文章。

17.9 问题

在LocusLink中选择一个感兴趣的人类基因，或直接输入“lipocalin”选一个基因，在NCBI、Ensembl和UCSC中查看该基因的各类性质，并列出一张表（如外显子/内含子结构、相关的EST数目、多态性、相邻的基因等）。看看这三个数据库得到的数据是否有不一致之处。然后，从包含这个基因的基因组区域截取一段DNA序列（约100000bp，FASTA格式），用Oak Ridge National Laboratory pipeline分析这段序列的特性并预测编码区（如图16-13～图16-16）（http://compbio.ornl.gov/tools/pipeline）。

17.10 自测题

1. 人类基因组大概有多大？

(a) 130Mb；

(b) 300Mb；

(c) 3000Mb；

(d) 30000Mb

2. 各种重复元件在人类基因组中大概占据了多少？

(a) 5%；

(b) 25%；

(c) 50%；

(d) 95%

3. 蛋白编码区占人类基因组的比例是多少？

(a) 1%～5%；

(b) 5%～10%；

(c) 10%～20%；

(d) 20%～40%

4. UCSC人类基因组浏览器和Ensembl、NCBI不同，因为

(a) 提供了大量的注释，其中一半是由外部用户贡献的；

(b) 提供了大量染色体图谱，包括保守同线区图谱；

(c) 提供了基于BLAST的基因组组装；

(d) 提供了结合公共测序联盟和Celera数据的组装

5. 人类基因组包含许多转座子来源的重复序列，它们被称为

(a) 死的化石；

(b) 年轻的活跃元件；

(c) 人类特异性元件；

(d) 反向重复

6. 片段性重复在人类基因组中占多少比例？

(a) <1%；

(b) 3%～5%；

(c) 20%～30%；

(d) 50%

7. 在人类基因组GC含量较高的那些区域中

(a) 基因密度倾向于比较低；

(b) 基因密度倾向于比较高；

(c) 基因密度多变；

(d) 基因倾向于含有更少的内含子

8. 和其他多细胞动物相比（如线虫、昆虫、小鼠）

（a）人类基因组有更多的蛋白编码基因；

（b）人类基因组有很多缺乏可识别的直系同源的特异基因；

（c）人类基因组有更高的GC含量；

（d）人类基因组有更多的多结构域蛋白、旁系同源基因和可变剪切

17.11 推荐读物

在本章中我们主要是根据人类基因组测序联盟的报告（IHGSC，2001）来叙述的。关于人类基因组，Celera公司的报告（Venter，2001）以及其他在“Science”、“Nature”中关于这个主题的文章也都值得一看。

首先完成测序的三个人类染色体是22号（Dunham等，1999）、21号（Hattori等，2000）和20号染色体（Deloukas等，2001）。这些重要的文章对于相应的染色体序列有着深入的分析。它们强调了完全测序对于获得更精确的注释和比较分析的重要性。

关于人类基因组的文章还有很多。2001年3月的“Genome Research”上Francis Collins、Eric Green和Aravinda Chakravarti的文章对于人类基因组计划目前的情况和前景发表了重要的看法。

如果是刚开始阅读人类基因组计划相关内容，可以从美国国家人类基因组研究所的网站开始：http://genome.gov。

参 考 文 献

Amir, R. E., et al. Rett syndrome is caused by mutations in X-linked *MECP2*, encoding methyl-CpG-binding protein 2. *Nat. Genet.* **23**, 185-188 (1999).

Avner, P., and Heard, E. X-chromosome inactivation: Counting, choice and initiation. *Nat. Rev. Genet.* **2**, 59-67 (2001).

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., and Eichler, E. E. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005-1017 (2001).

Baltimore, D. Our genome unveiled. *Nature* **409**, 814-816 (2001).

Bernardi, G. Misunderstandings about isochores. Part 1. *Gene* **276**, 3-13 (2001).

Crick, F. H. and Watson, J. D. Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).

Deloukas, P., et al. The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**, 865-871 (2001).

Dunham, I., et al. The DNA sequence of human chromosome 22. *Nature* **402**, 489-495 (1999).

Eichler, E. E. Segmental duplications: What's missing, misassigned, and misassembled—and should we care? *Genome Res.* **11**, 653-656 (2001).

Gardiner-Garden, M., and Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261-282 (1987).

Goodman, N. The IT GUY. *Genome Technol.* **1**, 55-59 (2002).

Green, E. D. Strategies for the systematic sequencing of complex genomes. *Natl. Rev. Genet.* **2**, 573-583 (2001).

Green, E. D., and Chakravarti, A. The human genome sequence expedition: Views from the "base camp." *Genome Res.* **11**, 645-651 (2001).

Haring, D., and Kypr, J. Mosaic structure of the DNA molecules of the human chromosomes 21 and 22. *Mol. Biol. Rep.* **28**, 9-17 (2001).

Hattori, M., et al. The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* **405**, 311-319 (2000).

Hubbard, T., et al. The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38-41 (2002).

IHGSC. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

Jurka, J. Repeats in genomic DNA: Mining and meaning. *Curr. Opin. Struct. Biol.* **8**, 333-337 (1998).

Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656-664 (2002).

Kent, W. J., et al. The human genome browser at UCSC. *Genome Res.* **12**, 996-1006 (2002).

Li, S., et al. Comparative analysis of human genome assemblies reveals genome-level differences. *Genomics* **80**, 138-139 (2002).

National Research Council. *Mapping and Sequencing the Human Genome.* National Academy Press, Washington, DC, 1988.

Ostertag, E. M., and Kazazian, H. H., Jr. Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* **11**, 2059-2065 (2001).

Pavlicek, A., Paces, J., Clay, O., and Bernardi, G. A compact view of isochores in the draft human genome sequence. *FEBS Lett.* **511**, 165-169 (2002).

Ponting, C. P. Plagiarized bacterial genes in the human book of life. *Trends Genet.* **17**, 235-237 (2001).

Ross, D. T., et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* **24**, 227-235 (2000).

Salzberg, S. L., White, O., Peterson, J., and Eisen, J. A. Microbial genes in the human genome: Lateral transfer or gene loss? *Science* **292**, 1903–1906 (2001).

Singer, C. *The Fasciculo Medicina Venice 1493.* R. Lier and Co., Florence, 1925.

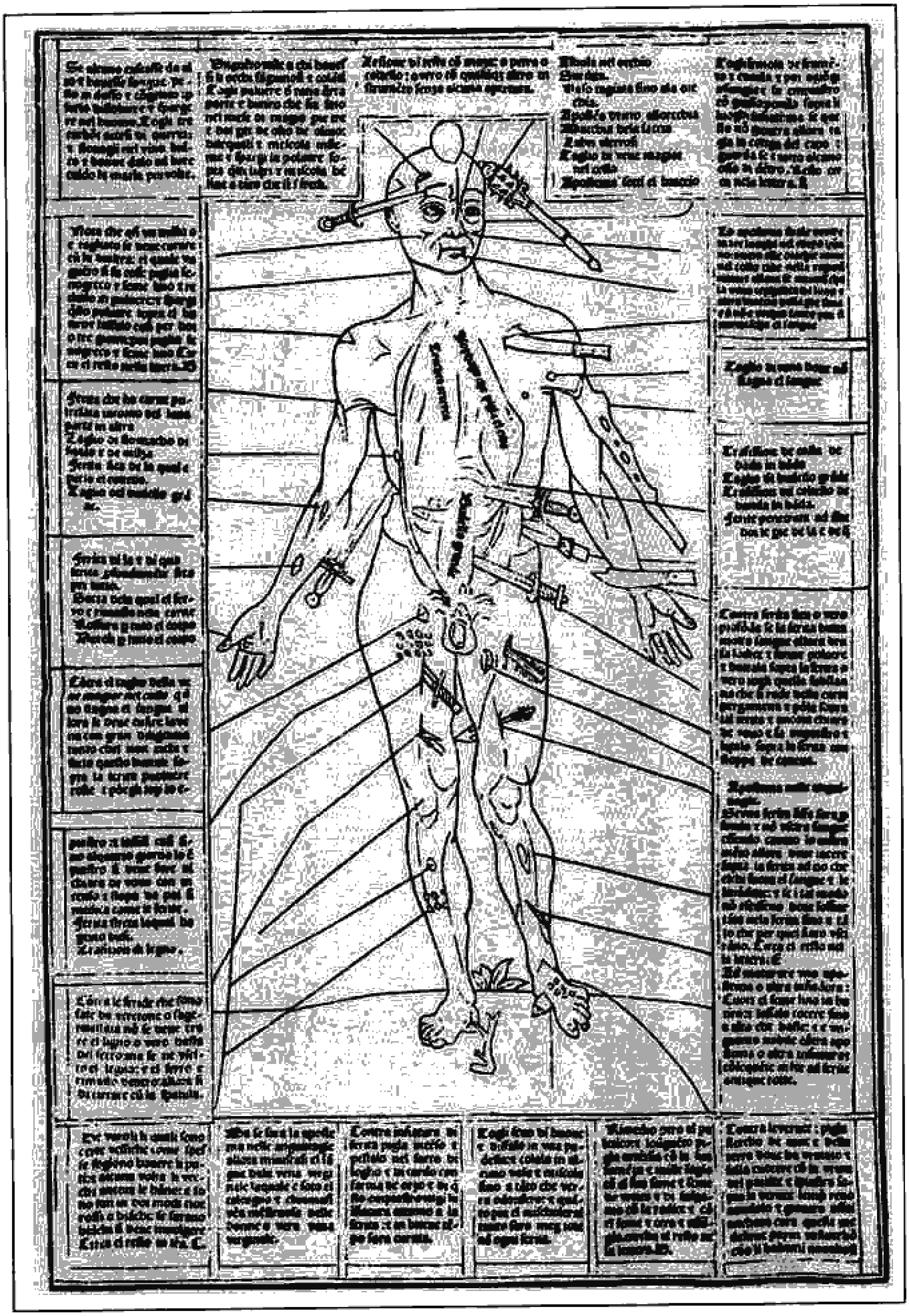
Toth, G., Gaspari, Z., and Jurka, J. Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Res.* **10**, 967–981 (2000).

Tycko, B., and Morison, I. M. Physiological functions of imprinted genes. *J. Cell. Physiol.* **192**, 245–258 (2002).

Venter, J. C., et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).

Wiemann, S., et al. Toward a catalog of human genes and proteins: Sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.* **11**, 422–435 (2001).

Yu, A., et al. Comparison of human genetic and sequence-based physical maps. *Nature* **409**, 951–953 (2001).



《医学论集》(Fasciculus Medicinae)(1491 年首次出版)，作者是 Johannes de Ketham，书中有一个“受伤的人”(如图)。这幅图叙述了许多疾病的发病过程和治疗疾病的处方。在当时，最普遍的医疗方法大概就是放血术。图中左下角的文字写道：“水痘，就是经常由小孩得，偶尔成人也会患的小疱；它有两种，一种是白色的，另一种是红色的，如果是白色的，应该把患者全身都包裹在亚麻布衣服里。”图中左上角的文字说的是“当人从高处摔下，他的血液会凝固，让这些血液在人体中扩散和溶解的方法是：取三块烧红的栎木炭加入上好的白酒中，让病人在每个早晨服用。”详见 Singer (1925)

第 18 章

人类疾病

生命是一种分子间的联系，而不是任何单个分子的性质，因此疾病才会危害生命。分子疾病是存在的，但是没有患病的分子。在分子水平上，我们仅仅发现了结构和物理化学性质上的变化。同样地，在分子水平我们几乎找不到任何好的准则在进化的尺度上把一个给定的分子放在“高”或“低”的位置上。人血红蛋白尽管在某些方面和马的不同（Braunitzer 和 Matsuda，1961），但是绝没有表现得更加高度的组织化。分子疾病和进化是属于生物体高级水平的表现。它们是紧密联系着的，之间没有明显的界限。分子疾病的机制代表着进化机制的一个部分。甚至主观地说，进化和疾病这两种现象有时会导致相同的结果。善恶观念的出现，通常被人们认为是从伊甸园中偷食善恶果的结果，但它也可能是进化所致的一种分子疾病。

——Emile Zuckerkandl 和 Linus Pauling（1962，第 189～第 190 页）

18.1 人类遗传疾病：DNA 变异的结果

DNA 序列上的变异是地球上生命的一个确定的特征。对每一个物种，遗传变异是为了适应进化而发生的 DNA 序列的相应改变，这成为进化的基础。进化就是物种适应它们生活环境的过程。当 DNA 的变化提高了一个物种的适应性，它的种群就能更成功地繁殖。当变化相对地不利于适应，这个物种可能会灭绝。对于一个物种的某个个体来说，一些突变提高了它的适应性，大部分突变对适应性没有影响，还有一些则降低了适应性（相对于正常的）。疾病可以定义为一个种群中对个体有害的不利于适应的改变，也可以定义为反常条件下物理化学功能的削弱。我们把注意力放在这些生理缺陷的分子基础上，如 DNA、RNA 和蛋白质。

人类疾病在本质上有着极大的多样性。这是如下一些原因造成的。

① 突变会影响人类基因组的每个部分。有着无限多的机会使得不利于适应环境的突变产生。这些可能是点突变，只影响单个核苷酸；也可能是大范围突变，影响一整条染色体甚至多条染色体。

② 突变导致疾病存在多种机制（总结在表 18-1 中）。基因功能被破坏有的是由于点突变而改变单个氨基酸残基而导致的；有的是由于 DNA 的删除或者插入，大小可以从一个核苷酸到一整条染色体，后者大于 100 万个碱基对（Mb）；有的是由于 DNA 片段方向的逆转。在很多情况下，同一个基因上不同类型的突变会导致不同的表现型。

表 18-1 遗传突变的机制

机 制	通常的效果	范 例
大范围突变		
删除		Duchenne 型肌营养不良症
插入		血友病
复制	无，基因破缺	Duchenne 型肌营养不良症
	剂量，基因完好	腓骨肌萎缩症
反转	—	血友病
扩展三联体	—	X 染色体易碎症
	获得新功能？	Huntington's 病
点突变		
沉默	无	囊性纤维化病
错义或框内删除	—，功能部分丧失，功能改变，良性	球蛋白
无意义	—	囊性纤维化病
框移动	—	囊性纤维化病
剪切(AG/GT)	—	球蛋白
剪切(外 AG/GT)	功能部分丧失	球蛋白
调控(TATA，其他)	功能部分丧失	球蛋白
调控(poly A 位点)	功能部分丧失	球蛋白

注：1. AG/GT 显示内含子中最初和最后两个规范的碱基对的突变，外 AG/GT 显示较不规范序列的突变。

2. 来源并改编自 Beaudet 等（2001），授权使用。

③ 几乎所有的基因都是通过产生蛋白质这样的基因产物来发挥功能。基因中的一个可导致疾病的突变会使基因产物没有正常的功能。这对于基因产物能正常表达并行使功能的细胞的能力有着深远的意义。

④ 个体和他（或她）所处环境的相互作用对于疾病的表现型有重要的影响。一般说来，同一对双胞胎可能会有完全不同的表现型。这些差异来自于环境的影响（或者说后天的效应）。同卵双生子之间的共同性表征着遗传和环境效应影响疾病的相对程度。甚至对于严重的遗传疾病（像自闭症和精神分裂症），这种疾病表征的相似度也永远不会达到 100%。

突变是 DNA 序列的改变，其原因也许是 DNA 复制或者修复过程中的错误，或是化学诱变剂的效应，或者是辐射导致。虽然突变这个概念与负面的含义相关，但是突变和固定性是进化的基本推动力。

从医学的观点来看，疾病是"机体的一种病理状态，表现出一系列的临床迹象和表征，实验室发现这种状态病并把它作为一种非正常实体而与正常状态或者其他病理状态区分开"（Thomas，1993）。紊乱是"精神或者机体的一种病理状态"（Thomas，1993）。综合征就是"由于解剖学上的、生理上的、生物化学上的特性造成的相互联系的功能之间的紊乱所表现出来的征兆。这些定义没有包含某个疾病的准确的成因，但是却提供了一个指导我们探索它的框架"（Thomas，1993）。

18.2 人类疾病的生物信息学展望

在第 1 章里，我们把生物信息学定义为使用计算机数据库和计算机算法来分析蛋白质、基因、基因组的一门学科。我们研究人类疾病的方法是还原论性质的，因为我们在寻找描述那些导致疾病的基因和基因产物的方法。然而，对疾病的分子基础的认识可能是和揭示整个人类的疾病的逻辑体系结合在一起的（Childs 和 Valle，2000）。当我们寻找生物信息学方法来研究人类疾病时，我们就会面临如何处理整个复杂生物系统的问题。即使我们找到了由于基因突变而导致人类疾病的基因，我们依然面临挑战，那就是如何把该基因的基因型和表型联系起来。我们现在根据各方面的综合信息能够做到的仅仅是确定单个基因的功能，以及该基因产物对细胞的功能（Childs 和 Valle，2000；Dipple 等，2001）。

生物信息学领域提供了帮助我们理解关于基因和环境对于疾病过程的每个方面的影响这样一些基本问题的方法来研究人类疾病。这个领域中一些影响我们对疾病认识的方法的例子将始终是本章的重点，包括如下部分。

① 就疾病的遗传基础是 DNA 序列变异的一种作用来说，DNA 数据库为我们提供了 DNA 序列比对的基本且必需的材料。这些数据库包括主要的、一般性的 DNA 序列库（如 GenBank/EBML/DDBJ，第 2 章），一般性的资源（如人类孟德尔遗传在线，OMIM），以及提供单个基因位点的序列变异数据的位点特异性数据库。

② 那些通过遗传连锁研究、群体研究或其他试验（后面讲到的）来寻找致病基因的遗传学家依靠物理和遗传图谱来尽力寻找突变的基因。

③ 当一个编码蛋白质的基因突变时，它会导致这个基因的蛋白质产物的三维结构发生变化。第 9 章中介绍的生物信息学工具可以允许我们预测蛋白质变体的结构，从这样的分析中我们可以推断出它功能上的变化。

④ 一旦一个突变的基因被确定下来，我们就想知道这种突变是怎样影响细胞功能的。在第 8 章中，我们已经介绍了多种方法来了解蛋白质的功能。在我们对啤酒酵母的讨论中，我们谈到了了解真核生物的蛋白功能的另外的高通量方法（第 15 章）。基因表达分析（第 6 章和第 7 章）在研究疾病状态的转录响应时被采用。

⑤ 我们或许能够通过确定人类一个特定基因在低等物种中的直系同源基因来获得对该基因功能的深入了解。我们将讨论在各种模式生物中找到的人类疾病基因的直系同源基因。

在本章中，我们首先介绍人类疾病的概况，包括疾病分类的方法。接下来，我们将在三个层次上考虑人类疾病的主题（图 18-1 给出要点）。首先，在分子水平上，我们将关注基因在疾病中扮演的角色。在讨论单基因（一个基因）无序性时，我们将介绍人类孟德尔遗传在线（OMIM），它是主要的疾病数据库。也有几百个位点特异的突变数据库是我们将要讨论的。我们还将考察和 RNA、蛋白质研究相关的生物信息学方法以及数据库。第二，我们将检索细胞和系统水平上的与疾病相关的网络资源，如细胞器疾病数据库。第三，我们考虑有机体水平上开发了哪些生物信息学工具用来刻画疾病的临床表征（如发作年龄、遗传模式、频率以及严重程度）？发展了哪些疾病的动物模型？我们将在模型生物中研究人类疾病基因的直系同源基因（如真菌和低等后生动物基因）。最后，我们将考虑那些已经被建立的提供人类疾病综合信息的数据库。

	水平	例子	生物信息学资源
分子水平	DNA		一般资源:OMIM,HMD 特定基因突变数据库
	RNA		基因表达数据库(微阵列) 基因表达数据库(SAGE)
	蛋白质		突变蛋白数据库(SwissProt) 二维电泳数据库
系统水平	组织		过氧化酶体病数据库 线粒体病数据库 溶酶体病数据库
	生物/系统		血液数据库 神经肌肉疾病数据库 心血管疾病数据库 肠胃系统疾病数数据库 视网膜疾病数据库
物种水平	临床表型	发作年龄 遗传模式 频率 严重性 组织牵连 畸形	
	动物模型	人类疾病基因	…鼠 …线虫 …蝇 …酵母 …其他
	组织和基金会		一般组织(NORD) 特定疾病组织

图 18-1 研究人类疾病的生物信息学资源在多个水平上被加以组织

18.3 Garrod 对疾病的观点

Archibald Garrod(1857～1936)对了解人类疾病的本质做出过重要的贡献。在 Garrod(1902)的论文中，他描述了他对尿黑酸尿症的研究，这是一种罕见的遗传疾病。在尿黑酸尿症中，尿黑酸-1,2-加双氧酶(HGD)失活或者缺少。这样导致的结果是氨基酸中的苯丙氨酸和酪氨酸不能正确地代谢，导致一种代谢物(尿黑酸)累积。这种代谢物在尿中被氧化后变成黑色。Garrod 从进化的角度考虑这种表征，注意到自然选择在化学过程中的影响。个体间代谢过程中的变化有可能会导致疾病的发生。

Garrod 洞察到在每个他所研究的罕见疾病中，病理表征反映出个体的化学特征。他进一步发现这种特征是固有的——他提出尿黑酸尿症以孟德尔隐性遗传方式传代。那时，人们认为几乎所有的疾病都是由外界因素引起的(如细菌感染)。在对于该病以及相关的隐性疾病(如胱氨酸尿症和白化病)的研究中，他提出这些疾病的表征是由于一个固有酶的失活或者化学差异导致(Scriver 和 Childs，1989)。他在他的第一本书——《先天的代谢缺陷》(1909)中描述了这种观点。Garrod 在 1923 年写道(Scriver 和 Childs 1989 年所引用)：

"如果一个物种中的某些个体在化学结构和化学行为上和这个物种中正常个体不同，那么这样的变体或者突变就显然具有被自然选择后保存下来的能力；现在不少生物学家把化学结构和功能指定为物种进化上的最重要的部分……极少数个体表现出显著背离正常代谢（就像卟啉尿症和胱氨酸尿症），但是我强烈地猜测那些被忽略的最小的背离是普遍的。另外，又如何解释疾病中由于遗传方面的原因而引起的部分呢？有一些疾病是从上一代传到下一代的……在孩童后期和成人早期哪些疾病倾向于发生……我们很难逃避一个结论，那就是虽然这些疾病并不是先天的，但是它们具有潜在的先天特性。"

Garrod 于是提出了一个新的关于先天因素怎样导致疾病的观点（在 Beadle 和 Tatum 提出一个基因编码一个蛋白质的假说之前），并且他从来没使用过"基因"这个词。但是我们现在知道他所描述的"先天特性"就是突变的基因。他工作的一个主要结论是由于遗传差异造成的化学特性是人类健康和疾病的主要决定性因素。虽然"化学特征"这个词现在不常用，但是这个概念在基因药理学领域有重要的意义。不是每一个人暴露在有感染性的媒介中都会生病（我们必须了解为什么会这样），也不是每个人吃药后的反应都是相似的。

Garrod 在第二本书——《疾病中的先天因素》（1931）中进一步发展了这些观点。这本书中他提出这样的问题：为什么某些个体容易受疾病影响——无论这种疾病是明显带有遗传性的或者是由于其他原因（如环境媒介）导致的。他认为化学特性使我们易受疾病感染。每个疾病的过程都受内在和外在因素的影响：我们的遗传互补以及所处的环境因素。在某些条件下（如代谢的先天缺陷），遗传因素扮演着更为重要的角色。在有传染性的疾病中，基因在决定个体的易感性以及机体对传染性媒质的响应方面也起着重要的作用。接下来我们将继续讨论各种各样的疾病。

18.4 疾病的种类

什么样的疾病使人痛苦？疾病应该怎样分类？我们首先把疾病分为 5 大类（图 18-2）。从根本上说，这些分类有些武断，因为每个疾病在不同的个体中会表现出很多表型，每个疾病也会和很多基因相关。

① 孟德尔疾病，也称作单个基因或者单基因疾病，主要是由于单个基因的突变而引起。在 Garrod（1909）的书中描述的孟德尔疾病在一般人群中出现得很罕见。总体看来，一般人群中孟德尔疾病的发生率约为 1%，并且主要出现在幼年时期（成年之前）。在这些疾病中，一个基因的突变对该疾病起着主要的影响。然而，即使是单基因疾病，许多其他的基因也可能会起到调节作用，从而影响疾病的表型。这样，在某个基因上有相同突变的两个个体可能会有完全不同的疾病表现模式（下文讨论）。

② 复杂疾病（如阿尔茨海默症和心血管疾病）是由于多个基因的缺陷引起的。这些疾病也称作多因素疾病，反映出它们既受遗传也受环境因素的影响。和单基因疾病相比，复杂疾病往往更加普遍。在美国，像心脏病、老年痴呆症、癌症以及糖尿病这样的慢性疾病是导致死亡和残疾的主要原因。这些疾病都有一定程度的遗传基础。

③ 从致死性染色体畸变在所有人类胚胎中的发生率达 50%～75%这点来看，染色体异常可以认为是最常见的导致疾病的原因。

④ 从出生到晚年，感染性疾病是全球人口死亡的首要原因。

疾病种类		每1000个中占有数
单基因疾病（孟德尔）		11
	常染色体显性	6
	常染色体隐性	3
	X-染色体隐性	1
	X-染色体连锁的智力障碍	1
复杂（多基因）疾病		630
	先天畸形	30
	中枢神经系统疾病	100
	心血管疾病	500
染色体疾病		高
传染性疾病		高
	病毒、细菌或其他病原体	
	例如艾滋病	
环境疾病		高
	外生因素，如毒素、光照	
	（非遗传因素）	
	例如铅中毒	

图 18-2 人类疾病可以按照病因来分类。它们包括单基因疾病（单个基因的突变：例子有苯丙酮尿症和镰刀型红细胞贫血）；复杂疾病（两个和多个基因的突变，如癌症和精神分裂症）；染色体疾病（如唐氏综合征）；传染性疾病；以及环境疾病。这些疾病的发病率数值只是一个近似的估计。总的来说复杂疾病比单基因疾病更为普遍。但是导致单基因疾病的遗传缺陷则更容易被发现。对于所有种类的疾病，病理生理学（即因疾病改变了的生理过程）受很多基因的影响

⑤ 环境疾病，被定义为由具有非遗传的原因所导致的疾病，也非常普遍。例如，感染性的疾病和环境毒素能导致与遗传突变无关的疾病。

有各种各样导致疾病的原因（遗传的或非遗传的）。与 Garrod 的观点一致，任何疾病的病理生理学都可能被认为是多基因的。两个受到相同致病原刺激的个体——不管致病原是病毒或者铅涂料或者是一个突变的基因——可能会有完全不同的反应。如一个人可能会生病，而另外一个则不会受到影响。对于任何疾病诱因条件的响应，遗传是一个很大的影响成分。

18.5 疾病的分类

我们在图 18-2 中描述了疾病的几种一般的类别。从生物信息学观点来看，我们感兴趣的是了解那些与疾病机制相关的基因组 DNA、基因，以及它们的基因产物。我们更感兴趣的是在整个进化中，突变对细胞功能以及对致病基因的比较基因组学的重要性。这种观点与

临床医生和流行病学家的观点有很大不同。

> 铅中毒是由环境因素导致疾病的一个例子。根据美国疾病控制和预防中心的报道，美国大约有 8%的儿童血液含铅达到警戒水平。见 http://www.cdc.gov/mmwr/preview/mmwrhtml/mm4950a3.htm#tabl。
>
> 病理学是研究疾病的成因和本质的一门学科。病理生理学是研究疾病怎样改变正常生理过程的一门学科。
>
> 死亡因素排名某种意义上是个随意的过程，它依赖于因素列表的选择和统一排名程序的使用（Anderson，2001）。比如表 18-2 列出的不包括肺癌（肿瘤的一个子集）或者机动车辆事故（事故的子集），不然它们任何一个都将排在前 10 位中。

有许多对疾病分类的方法。一种是描述死亡率统计数字。这些数据（基于美国的死亡证明书）包括导致死亡的原因排序（表 18-2）（Aderson，2001）。这些信息有助于确定最常见的疾病。

表 18-2 死亡主要原因（美国，1999 年）

排名	死亡原因	死亡人数/人	占总死亡数的百分比/%
1	心脏病	725192	30.3
2	恶性肿瘤	549838	23.0
3	脑血管疾病	167366	7.0
4	慢性减弱呼吸疾病	124181	5.2
5	事故	97860	4.1
6	糖尿病	68399	2.9
7	流行性感冒和肺炎	63730	2.7
8	阿尔茨海默症	44536	1.9
9	肾炎，肾综合征，肾病	35525	1.5
10	败血症	30680	1.3
	其他所有原因	484092	20.2
	所有原因	2391399	100.0

另一个划分人类疾病的方法是根据患病人群的百分比（图 18-3）或者根据伤残调整生命年（DALYs）（图 18-4）来衡量全球疾病方面的负担。在全球，像抑郁症和心脏病这样的非传染性疾病迅速取代了传染性疾病以及营养不良而成为导致残疾和过早死亡的首要原因（Murray 和 Lopez，1996）。

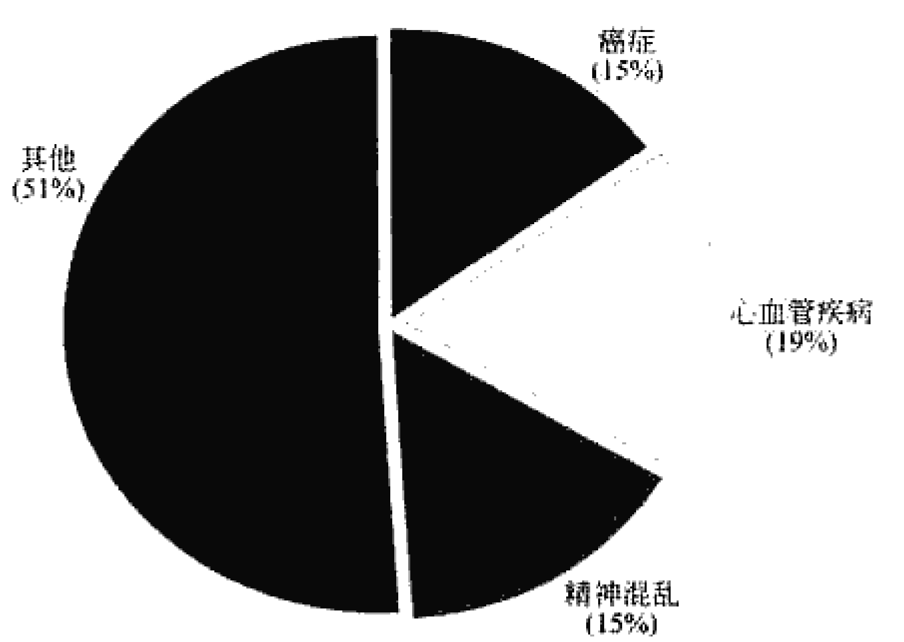


图 18-3 全球疾病负担。改编自 http://www.surgeongeneral.gov/library/mentalhealth/chapter6/sec2.html#figure6.3

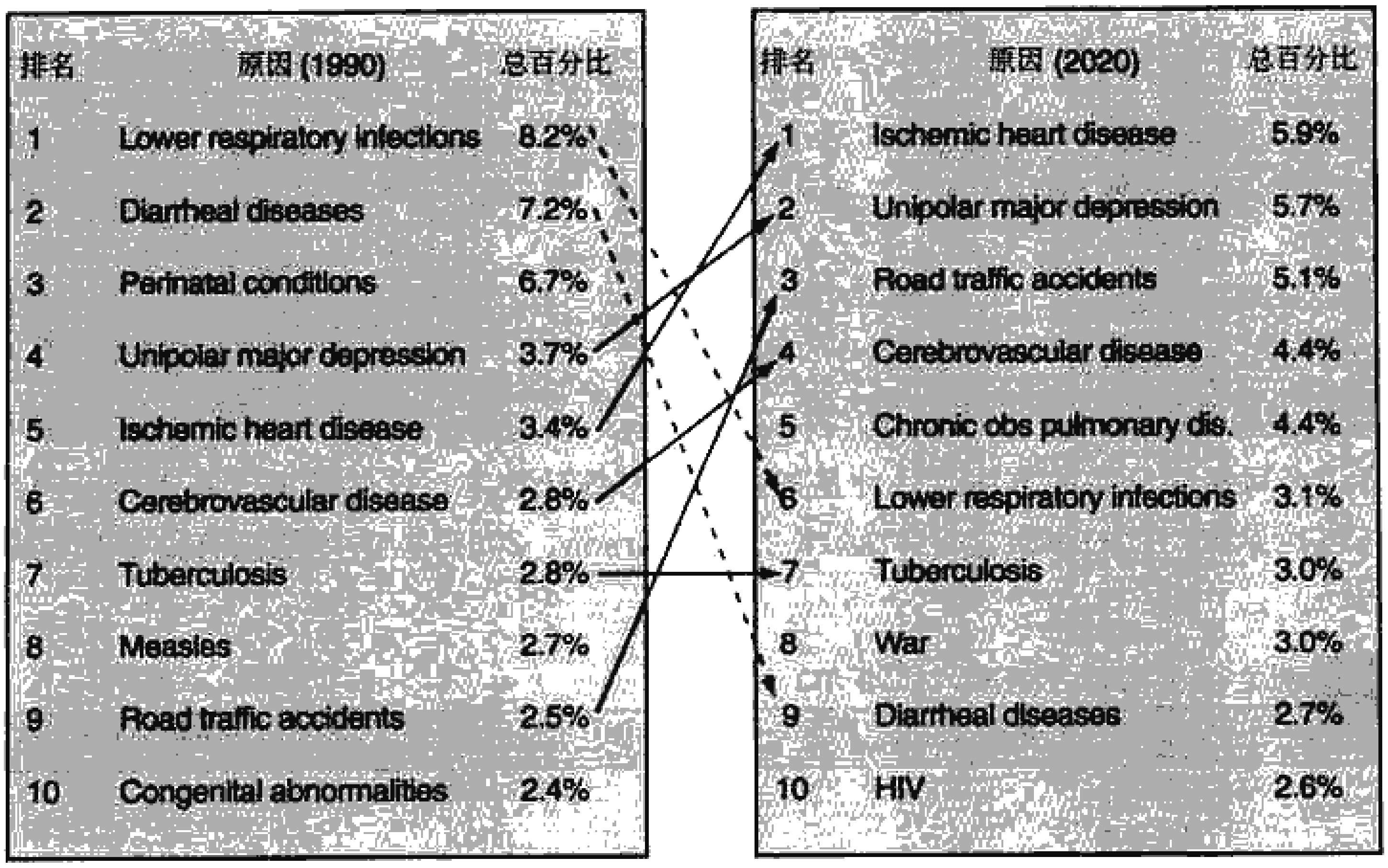


图 18-4 以失能调整生命年（DALYs）来衡量1990年和2020年（计划）的全球疾病负担。在女性中和在发展中国家，疾病负担首要因素的变化反映出单极的衰退。来自 Murray 和 Lopez（1996，第375页），授权使用

国际疾病与有关健康问题统计分类组织（ICD）提供了极为详尽的发病率数据列表。这项资源是世界卫生组织提供的，用于疾病的分类（表18-3）。

表 18-3 ICD 分类系统

1. 传染和寄生疾病
2. 肿瘤
3. 内分泌、营养和代谢疾病，以及免疫失调
4. 血液和造血系统疾病
5. 精神疾病
6. 神经系统和感觉器官疾病
7. 循环系统疾病
8. 呼吸系统疾病
9. 消化系统疾病
10. 泌尿生殖系统疾病
11. 妊娠、分娩和产后期并发症
12. 皮肤和皮下组织疾病
13. 肌肉、骨骼以及结缔组织疾病
14. 先天异常
15. 源于围产期的特定疾病
16. 症状、征候以及诊断不明疾病
17. 创伤及中毒

死亡率统计表列出了最为普遍的疾病。我们感兴趣的是全部的疾病分布谱，包括罕见的疾病。罕见疾病被定义为患病人数小于200000人的疾病。在美国，估计有2500万人（几乎占总人口的10%）忍受6000种罕见疾病中的一种或多种造成的痛苦。

附：NIH 的疾病分类：MeSH 术语

NLM（美国国家医学图书馆）开发了医学主题标题（MeSH）术语作为一种统一的语言用于生物医学文献数据库的搜索。2003 年的 MeSH 术语系统包含了 23 个疾病种类（图 18-5）。NCBI（美国国立生物技术信息中心）的 PubMed 数据库使用的也是这种分类系统。

1. ⊞ **Anatomy [A]**
2. ⊞ **Organisms [B]**
3. ⊟ **Diseases [C]**
   - Bacterial Infections and Mycoses [C01] +
   - Virus Diseases [C02] +
   - Parasitic Diseases [C03] +
   - Neoplasms [C04] +
   - Musculoskeletal Diseases [C05] +
   - Digestive System Diseases [C06] +
   - Stomatognathic Diseases [C07] +
   - Respiratory Tract Diseases [C08] +
   - Otorhinolaryngologic Diseases [C09] +
   - Nervous System Diseases [C10] +
   - Eye Diseases [C11] +
   - Urologic and Male Genital Diseases [C12] +
   - Female Genital Diseases and Pregnancy Complications [C13] +
   - Cardiovascular Diseases [C14] +
   - Hemic and Lymphatic Diseases [C15] +
   - Neonatal Diseases and Abnormalities [C16] +
   - Skin and Connective Tissue Diseases [C17] +
   - Nutritional and Metabolic Diseases [C18] +
   - Endocrine Diseases [C19] +
   - Immunologic Diseases [C20] +
   - Disorders of Environmental Origin [C21] +
   - Animal Diseases [C22] +
   - Pathological Conditions, Signs and Symptoms [C23] +
4. ⊞ **Chemicals and Drugs [D]**
5. ⊞ **Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]**
6. ⊞ **Psychiatry and Psychology [F]**
7. ⊞ **Biological Sciences [G]**
8. ⊞ **Physical Sciences [H]**
9. ⊞ **Anthropology, Education, Sociology and Social Phenomena [I]**
10. ⊞ **Technology and Food and Beverages [J]**
11. ⊞ **Humanities [K]**
12. ⊞ **Information Science [L]**
13. ⊞ **Persons [M]**
14. ⊞ **Health Care [N]**
15. ⊞ **Geographic Locations [Z]**

图 18-5 美国国家医学图书馆中的医学主题向导（MeSH）条目系统包含 15 个主要类目（2003 年版本）。疾病类目进一步包含 23 条向导。见 http://www.nlm.nih.gov/mesh/

MeSH 术语是一个受控的词汇表用来索引 MEDLINE（以及基于 MEDLINE 的 PubMed）。在 NLM MeSH 站点搜索术语“Rett Syndrome”将会显示 MeSH 中的有层次的树状结构：Rett Syndrome 被分别列在智力缺陷、神经退行性疾病、先天性遗传疾病这些类目下。

18.6 单基因疾病

我们在疾病的分子本质上的观点在最近的十多年有了发展。以前，遗传学家认识到简单特性和复杂特性之间的两面性。最近渐渐认识到所有的特性都是一个连续统一体的一部分，单个特性遵从孟德尔法则遗传。表 18-4 中列出了几种单基因疾病。作为单基因疾病的一个

例子，我们考查镰刀型红细胞贫血症（框18-1）。1949年Linus Pauling和他的同事描述了镰刀红细胞血红蛋白的反常电泳行为（Pauling等，1949）。随后发现一个氨基酸的替代导致了镰刀红细胞的反常行为，这也是镰刀红细胞贫血的基础。这是一个在常染色体上以隐性方式遗传的单基因疾病。但基因疾病在一般人群中往往是罕见的。注意到镰刀型红细胞贫血症是由血红蛋白中的氨基酸有一个发生特异性突变导致的结果。虽然镰刀型红细胞疾病有共同的特征（如镰刀状的红细胞），但这不是一个单基因疾病表型。这种多效性表型是由于其他基因的影响而造成的。

表18-4 单基因疾病的例子

机 制	疾 病	发生频率
常染色体显性	BRCA1乳腺癌及BRCA2乳腺癌	1/1000(犹太人为1/100)
	亨廷顿舞蹈病	1/2500
	Ⅰ型神经纤维瘤	1/3000
	结节性脑硬化	1/15000
常染色体隐性	白化病	1/10000
	镰刀型红细胞贫血症	1/655(美籍非洲人)
	囊肿性纤维化	1/2500(欧洲人)
	苯丙酮尿症	1/12000
X染色体连锁	A型血友病	1/10000(男性)
	葡萄糖-6-磷酸脱氢酶失活	可变(男性可到1/10)
	X染色体易碎综合征	1/1250(男性)
	色盲	1/12(男性)
	蕾特症	1/20000(女性)
	大脑白质硬化症	1/17000

注：改编自Beauder等（2001，第9页），授权使用。

框18-1 镰刀型红细胞贫血症

我们的细胞依靠氧气生存，血液把氧气运送到全身。然而，氧气是一种不溶于水的分子，需要载体蛋白——血红蛋白在血液中运输（肌肉细胞中运输氧气的同源蛋白是肌红蛋白）。成熟的血红蛋白是由两条α链和两条β链构成。染色体11p15.5的β链（NM_000518和NP_000509）上的突变导致镰刀型红细胞贫血症。病人的红细胞可以想像成弯曲的镰刀状，这种外形反映出血红蛋白在氧气浓度较低时的聚集。

镰刀型红细胞贫血症是美国最常见的遗传性血液疾病，500名美籍非洲人中就有1人患病。它是一种常染色体隐性遗传疾病。杂合子（具有一个正常血红蛋白β基因拷贝和一个突变拷贝的个体）具有一些抵抗疟原虫（*Plasmodium falciparum*）的能力。这可能是因为正常红细胞感染疟原虫后被破坏的缘故。这样，由于自然选择的进化压力保存了受疟原虫威胁的人群中的HBS突变。

下面是关于镰刀型红细胞贫血症的网络资源：

资 源	链 接
NIH事实表单	http://www.nhlbi.nih.gov/health/public/blood/sickle/sca_fact.pdf
基因和疾病(NCBI)	http://www.ncbi.nlm.nih.gov/disease/sickle.html
美国镰刀细胞疾病协会	http://www.sicklecelldisease.org

蕾特症（Rett Syndrome）是单基因疾病的另一个例子（框18-2）。这种疾病几乎都发生在小女孩中。患有蕾特症的女孩虽然出生时表面上看起来很健康，但是从出生后6~18个月

开始就会出现一系列征兆。她们失去了手的有目的的运动能力，而且表现出典型的手部扭曲行为。她们所掌握的任何语言能力消失了，可能会表现出类似于孤独症的行为。蕾特症是由编码 MeCP2（一种结合到甲基化的 CpG 岛上的转录阻遏蛋白）的基因上的突变引起的（Amir 等，1999）（见第 17 章）。现在还不知道为什么一个在整个身体中行使功能的转录阻遏蛋白的突变能导致机能性神经系统疾病。

我们在第 9 章中研究了正常 β 球蛋白（HBB）的结构以及最常见的变体形式（HBS）。E6V 的替换（第 6 位上的谷氨酸被替换成缬氨酸）给蛋白增加了一个磷酸化位点，促进了球蛋白分子的凝聚，形成了镰刀状的红细胞。

你可以在线阅读 Pauling 等的文章（1949）：http://profiles.nlm.nih.gov/MM/B/B/R/L。

美国国家医学图书馆（NLM）提供了一个科学站点（http://profiles.nlm.nih.gov），可在线获得几位著名生物学家的所有出版物。这些科学家包括 Linus Pauling 和其他荣获诺贝尔奖的人，如 Barbara McClintock、Julius Axelrod、Oswald Avery。

镰刀型红细胞贫血症作为一种单基因疾病来说非常普遍，这大概是由于它能提供对杂合子的保护以抵抗疟原虫的缘故（框 18-1）。

框 18-2 蕾特症

蕾特症（RTT：遗传编号 312750）是一种主要发生在女性中的发育性神经综合征（Hagberg 等，1983；Armstrong，1997；Naidu，1997）。患病的女性在出生后一个月内表面上发育正常，但接下来就会出现发育停滞。它还伴随着大脑生长迟缓、丧失语言和交际能力、严重的智力迟钝、躯体运动失调，以及特征性的手部扭曲运动等症状。显著的神经病理学特征包括多处脑皮质皮层厚度减少、神经元细胞体积减小、树突分支显著减少（Jellinger 等，1988；Bauman 等，1995；Belichenko 等，1997）。

Xq28 上的甲基化 CpG 结合蛋白 2（MECP2）基因的突变在多例蕾特症中被发现（Amir 等，1999；Wan 等，2000；Amir 等，2000；Bienvnu 等，2000；Cheadle 等，2000；Huppke 等，2000；Kim 和 Cook，2000；Van den Veyver 和 Zoghbi，2000；Xiang 等，2000）。*MeCP2* 在整个基因组中结合到甲基化的 CpG 二核苷酸上，并在辅助抑制物 mSin3A 以及染色质重构组蛋白去乙酰基酶 HDA1 和 HDA2 的协助下参与甲基化依赖的基因表达抑制。*MeCP2* 的 mRNA 在多种组织中的表达情况以及它在多条染色体上和 DNA 转录调控元件中的作用表明 *MeCP2* 是一个全局基因表达抑制物（Nan 等，1997）。DNA 甲基化依赖的基因表达抑制与遗传印记、X 染色体失活、致癌作用以及组织特异性的基因表达相关联（Cedar，1988；Allaman-Pillet 等，1998；Razin，1998；Ng 和 Bird，1999）。

为什么有些组织能不受 *MeCP2* 突变的影响？这可能是由于组织特异性的基因功能冗余，或者有其他的补偿机制。这为生物信息学工具怎样用来研究人类疾病的许多不同方面提供了一个例子。

虽然蕾特症是由单个基因的突变所引起，但是它也证明了人类疾病不同寻常的复杂性，即使是单基因疾病。

① 这种疾病主要发生在女性。这可以被解释为 *MeCP2* 基因位于 X 染色体上：这个基因上的突变可能在胚胎时期对男性是致死的（仅有一条 X 染色体），而女性则会有疾病的表征是因为她们的基因中有一个正常的和一个突变的拷贝。相反，更可能的解释是绝大多数突变发生在父亲身上，即父亲是健康的，但是一个精子出现新的突变并且遗传给女儿。这样，所有的儿子（XY）从父亲那里获得一条正常的 Y 染色体，而女儿则从父亲那里获得一条突变的 X 染色体。

② 发现 *MeCP2* 突变可以导致蕾特症以后，又发现一些智力迟钝的男性的这个基因也发生突变（Hammer 等，2002；Geerdink 等，2002；Zeev 等，2002）。但是，突变在男性中的

表征与在女性中完全不同，常常出现严重的新生儿脑病。男性只有一条X染色体，突变的基因实际上将会对身体的每个细胞产生不利的影响。相反，女性则遭受偶然的X染色体失活：有两条X染色体拷贝，每个细胞只表达其中的一条（或者是母亲的，或者是父亲的，发育早期随机选取）。女性的X染色体等位基因在表达上是一个嵌合体，一个患蕾特症的女性整个身体平均只有50%的正常细胞。

③ 虽然蕾特症是由一个编码转录阻遏蛋白的基因突变引起的，但是几乎可以确定这种突变的结果关联到其他许多基因的表达后继效果。这样，与其他单基因疾病相似，许多其他的基因受到关联而影响到该病的表征。

④ 两个在 *MeCP2* 基因上具有相同突变的女性可能有完全不同的疾病表征（视病情的严重程度）。就像在其他许多单基因疾病中看到的一样，对这种观察结果有两种主要的解释。a. 调节基因可能会影响疾病的进程（Dipple 和 McCabe，2002）。已经在镰刀型红细胞贫血症、肾上腺脑白质营养不良症、囊性纤维化病以及巨结肠病的病人体内发现了调节基因。很显然绝大多数（如果不是全部）单基因疾病是复杂的。b. 各种后生性的因素可能显著地影响临床表征。例如，基因组DNA的甲基化状态决定了 *MeCP2* 突变的分子结果。X染色体失活有时候是扭曲的，这样疾病的表征就更加严重（如果 *MeCP2* 突变的X染色体拷贝优先表达）或者更轻（如果正常的X染色体拷贝被选择表达）。

18.6.1 OMIM：有关人类疾病的重要生物信息学资源

1966年Victor A. McKusick创办了人类孟德尔遗传MIM。它的在线版本OMIM于1995年和NCBI整合。可以通过网址 http://www.ncbi.nlm.nih.gov/omim 或者通过NCBI的主页访问。

考虑孟德尔疾病时，我们现在转向OMIM，这是一个全面的人类基因和遗传疾病的数据库（Hamosh等，2000，2002）。OMIM数据库包含超过12000个人类疾病以及相关基因的目录条目。OMIM的焦点是先天性遗传疾病。OMIM的名称表明，该数据库与孟德尔遗传学相关。这是世代间信息传递的遗传特点。关于复杂疾病（见下文）中的遗传突变了解得相对较少，数据库中也没有包含染色体疾病（见下文）。这样它的焦点就是对单基因疾病的一个全面的考察。

我们用蕾特症和 *MeCP2* 分别作为疾病和疾病相关基因的例子来查询OMIM。OMIM可以从NCBI的站点上搜索到，并且通过LocusLink链接。在OMIM站点内，有一个搜索页面允许你查询各种主题，包括染色体、图谱位置，或者临床信息（图18-6）。结果包含蕾特症以及相关基因 *MeCP2*（图18-7）。由于 *MeCP2* 的突变和不同的表征相联系，所以它们也被列出来（如智力缺陷，该病是X染色体连锁的）。另一个相关的疾病被列出来，即安琪曼症，因为一些 *MeCP2* 突变的病人具有这种病的临床特征。

接着，用OMIM的标识符#312750我们可以查看蕾特症的条目（图18-8）。OMIM中的每一个条目都与一个编号系统相联系。6位代码的第一位表示相关联基因的遗传模式（表18-5）。蕾特症条目前面标记的数字表示这种表征可以由不只一个的基因突变所引起。第一个数字（3）表示这种疾病是X染色体连锁的。

MeCP2（*300005）的OMIM条目包含一个星号，星号表示该基因和其他加星号的基因一样具有独特的表型，且它们的遗传模式已被证实。条目包含像蕾特症的动物模型信息这样的目录数据。

OMIM链接到一个基因图上，它提供了一个有关疾病位点的细胞遗传位置列表（图18-9）。这个基因图进一步链接到NCBI的图浏览器（图18-9，箭头1）以及鼠类同源基因的

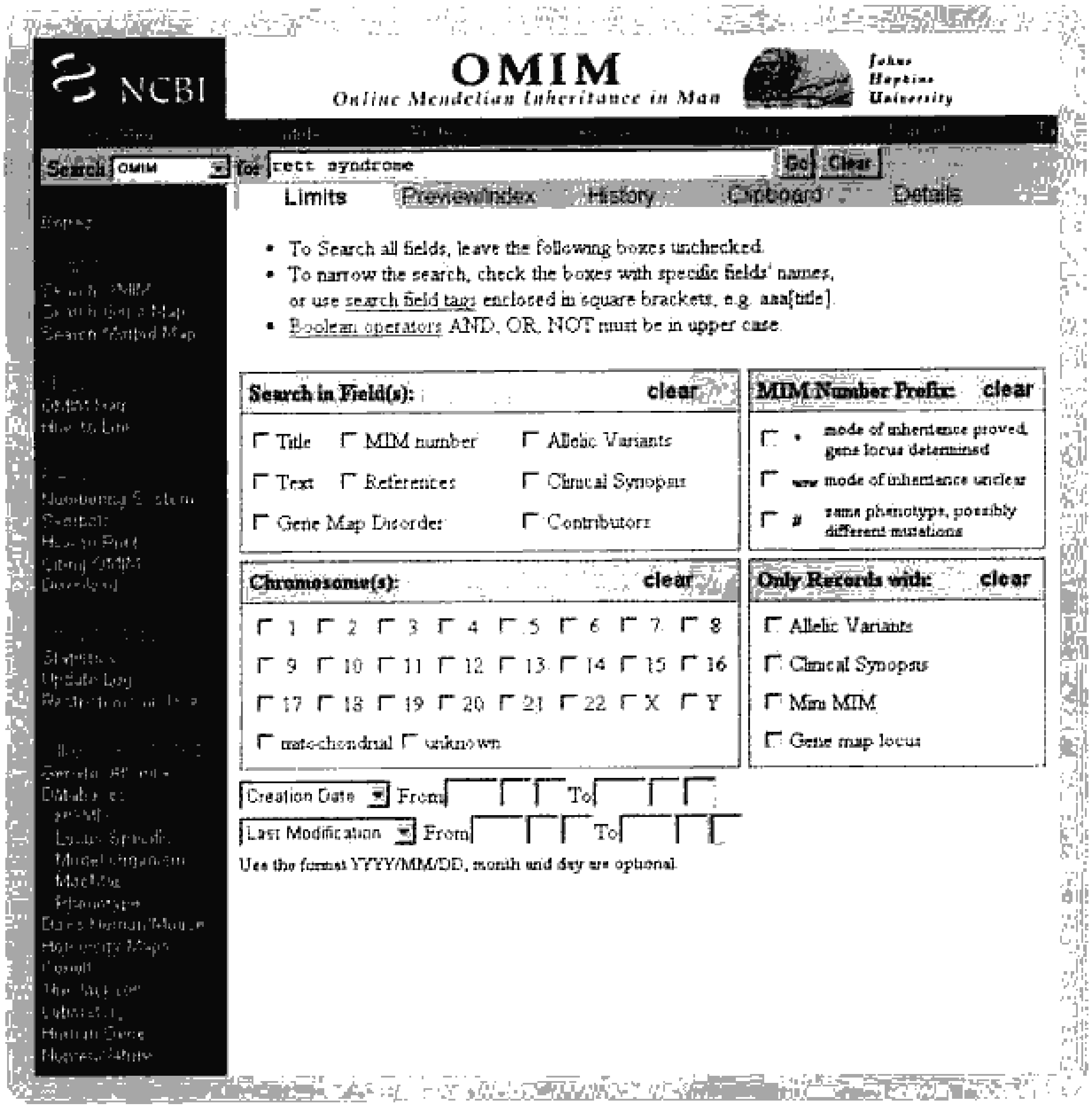


图 18-6 人类孟德尔遗传在线（OMIM），可通过 NCBI 网站访问（http://www.ncbi.nlm.nih.gov/omim），可按照作者、基因标识或染色体等标准进行文本搜索

表 18-5 OMIM 编号系统

OMIM 编号	表 型	OMIM 标识符	疾病(例子)	染 色 体
1——	常染色体显性	* 143100	亨廷顿病	4p16.3
2——	常染色体隐性	* 209850	孤独症，对其具易感性(AUTS1)	7q
3——	X 染色体连锁的位点或表型	# 312750	雷特综合征	Xq28
4——	Y 染色体连锁的位点或表型	* 480000	性别决定区	Yp11.3
5——	线粒体染色体的位点或表型	# 556500	帕金森病	—
6——	常染色体的位点或表型	# 603903	镰刀型红细胞贫血症	—

注：1. 条目前带“*”表示由该位点的基因确定的表型与其他的是不同的，而且这种表型的遗传模式已经被证实(根据作者和编辑的审查)。通常，每个基因位点仅有一个加 * 的条目。条目编号前没有 * 表示其遗传模式没有被证实，虽然已有推测；或者这个基因与其他条目所包含的基因位点的分离情况还不清楚。符号“#”表示这种表型可以由两个或多个基因中的一个发生突变而引起。对于 AUTS1 条目，数字 1 表示这是几个孤独症怀疑基因位点中的第一个。

2. 以 1 和 2 开始的条目表示的是 1994 年 5 月之前的数据；而那些以 6 开始的条目是 1994 年 5 月之后创建的。

资源上。OMIM 病理图也提供了细胞遗传位点，但它是按字母顺序组织的。OMIM 现有的数据在表 18-6 中以染色体的形式列出。

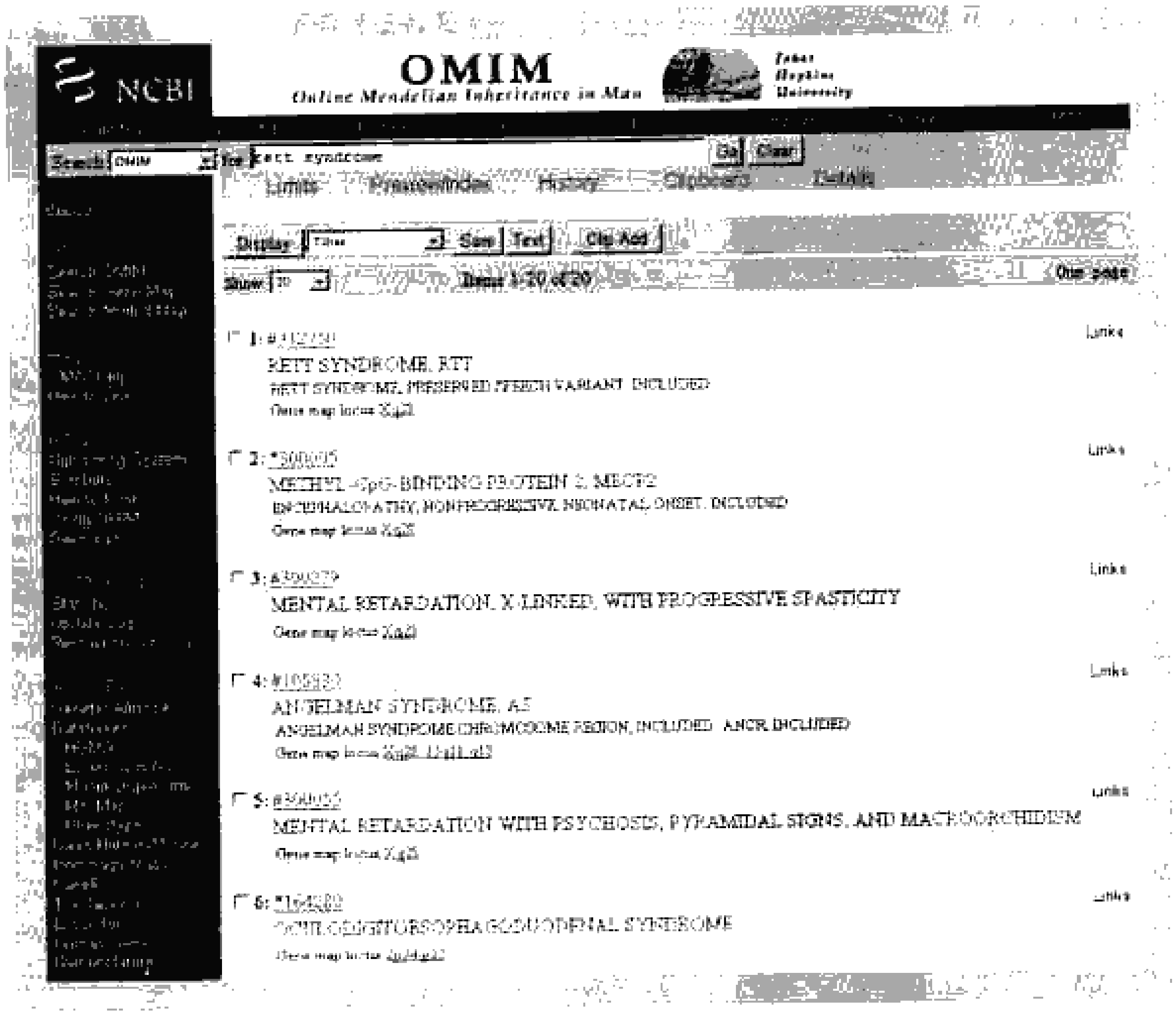


图 18-7 在 OMIM 中搜索“蕾特症”会产生该病症条目、与 *MeCP2*（该基因突变导致蕾特症）相关的条目、相同基因突变导致智力障碍（如 X 染色体连锁的伴随进行型痉挛的智力障碍）的条目以及和其他有着类似临床症状的疾病的区别（如安格曼症）

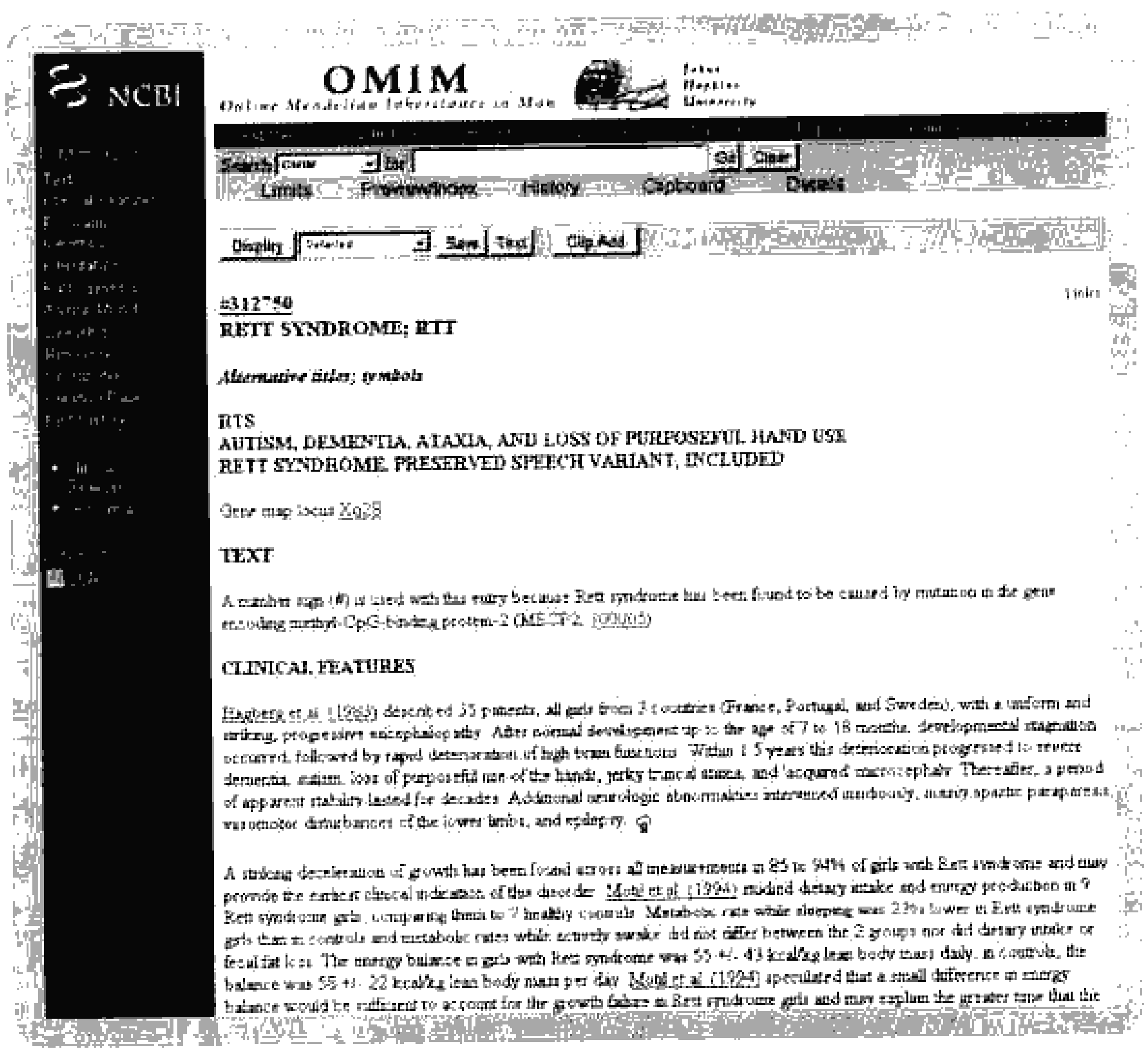


图 18-8 OMIM 的蕾特症条目包含其 OMIM 编号（# 312750），旁边列有多种信息，如对临床特点和现有动物模型的描述

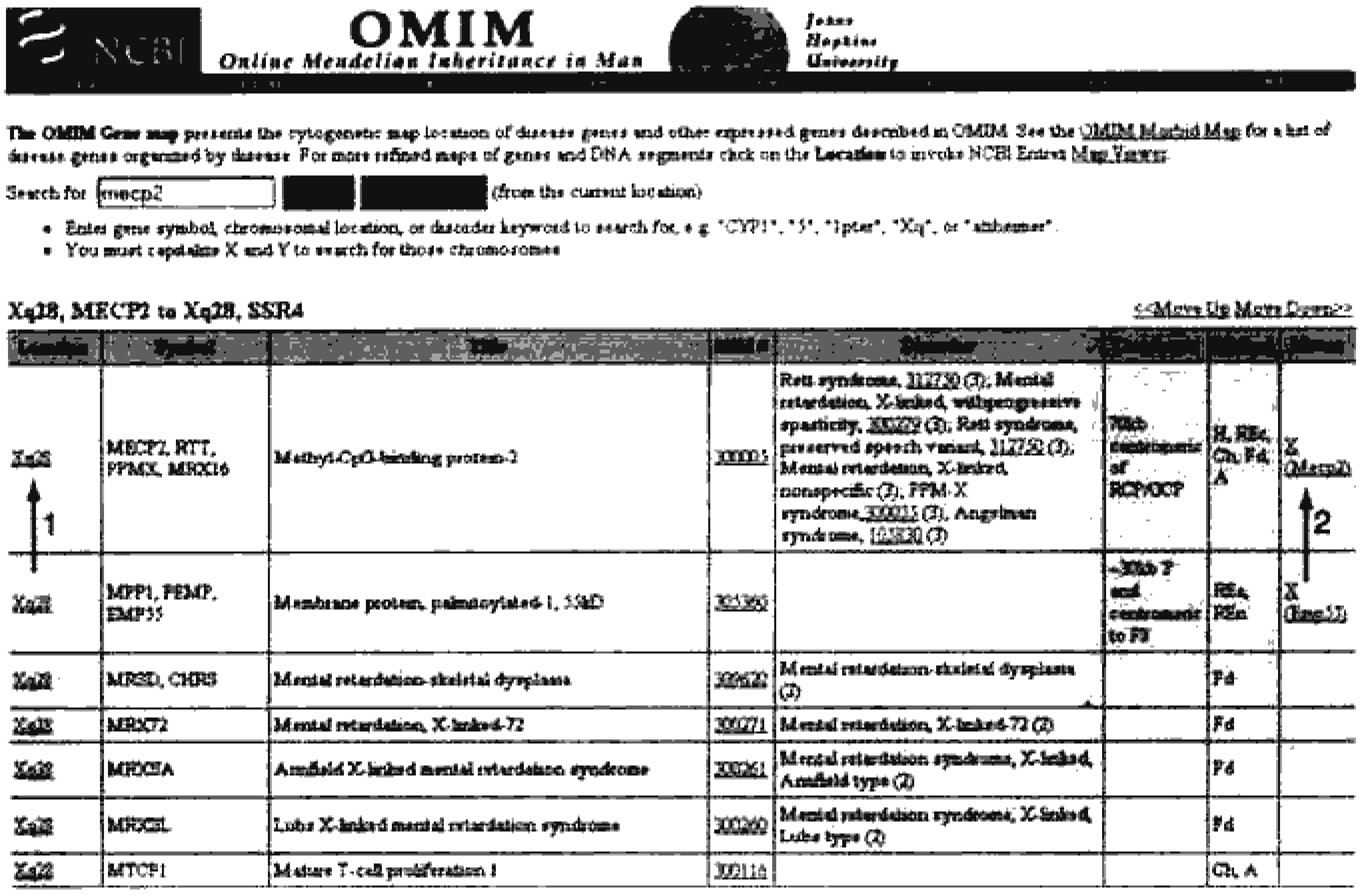


图 18-9 OMIM 的雷特症条目链接到 OMIM 的基因图谱上。这张图提供了一个列表，列出了疾病的细胞遗传位点，点击位置链接（箭头 1）转到条目的图谱查看器（图 17-2）。点击关于鼠信息的链接（箭头 2）转到鼠基因组信息站点的 *MeCP2* 页面（http://www.informatics.jax.org）

表 18-6 OMIM 中人类在每条染色体上的基因大纲（2002 年 11 月）

染色体号	基因位点	染色体号	基因位点	染色体号	基因位点
1	795	9	294	17	474
2	515	10	271	18	120
3	432	11	516	19	537
4	305	12	421	20	185
5	413	13	138	21	114
6	484	14	252	22	189
7	382	15	234	X	493
8	287	16	308	Y	34

注：1. 基因位点总数为 8193。

2. 参见 http://www.ncbi.nlm.nih.gov/entrez/Omim/mimstats.html。

OMIM 条目的一个重要特征是条目都包含一个等位基因变异列表。其中绝大多数代表的是导致疾病的变异。几个 *MeCP2* 等位基因变异条目的一个例子如图 18-10 所示。这些等位基因变异提供了对已知的人类所有致病的突变基因的一个大致了解。

OMIM 现有的数据总结在表 18-7 中。OMIM 一直是有关人类基因组信息的重要而全面的资源库。数据库由专门的管理者维护和更新（Hamosh 等，2002）。许多其他的数据库加入了 OMIM 标识符，从而对疾病相关基因提供一个通用的说明。

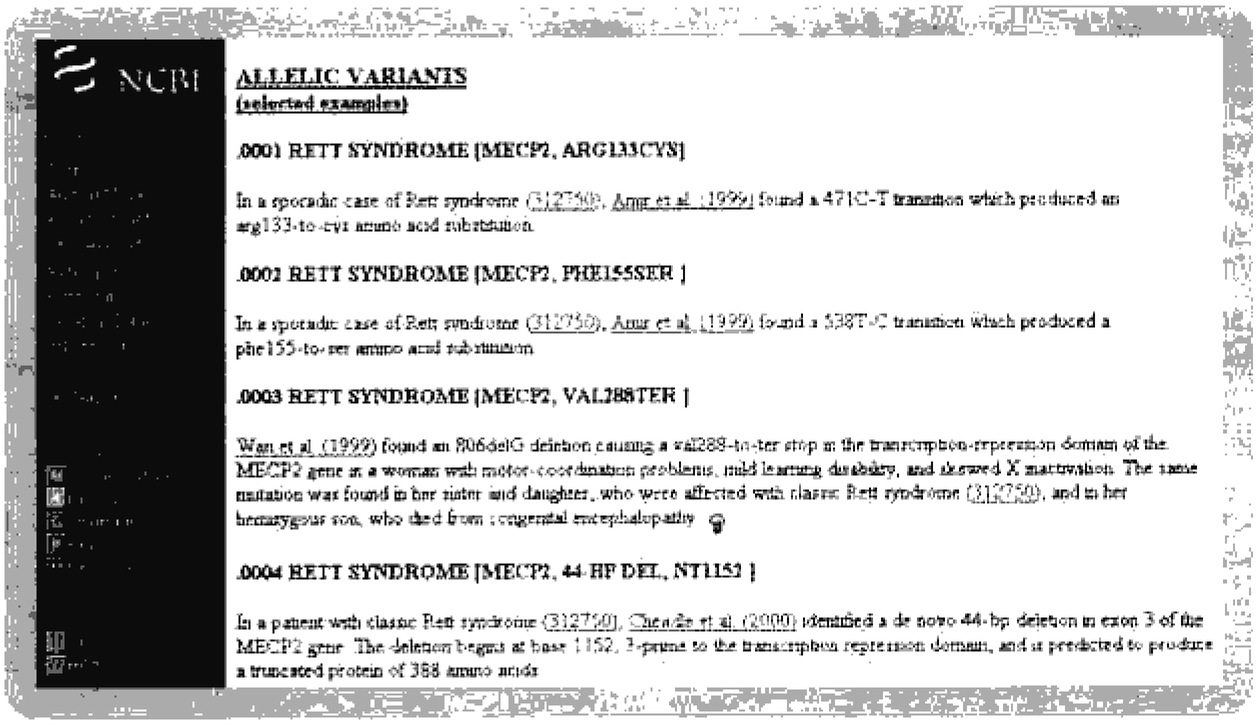


图 18-10 OMIM 的 *MeCP2* 条目包含等位突变，其中大多数是致病突变。一些等位突变表现出和疾病无关的多态性

表 18-7 OMIM 现有的数据情况（2002 年 11 月）

项　目	常染色体	X 连锁	Y 连锁	线粒体	总　计
已经确定的基因或表型位点(＊)	9792	538	39	37	10406
表型描述(＃)	1065	90	0	23	1178
其他位点或表型(无前缀)	2263	158	2	0	2422
总计	13120	786	41	60	14017

注：1. ＊和＃号的含义同表 18-5。

2. 参见 http://www.ncbi.nlm.gov/entrez/Omim/mimstats.html。

18.6.2 其他重要的人类疾病数据库

其他数据库提供了对人类疾病基因的全面了解，例子之一是 GeneCards（Safran 等，2002）。这是一个人类基因概要数据库，包含了关于人类疾病基因的丰富信息。GeneCards

> 等位基因变体的选择基于以下标准：首次被发现的突变具有高的种群频率，或者具有不寻常的致病机制。OMIM 中的一些等位基因变体表现出多态性，如果它们和常见的疾病表现出正相关则可能有特殊的意义（见下文）。
>
> 想要找到和人类疾病相关的一组基因，可以去 LocusLink 搜索，限制条件为人类，然后输入问题“disease_known AND has_seq.”现在有关于人类遗传疾病并且序列已知的条目有 1521 条（2003 年 6 月）。

和 OMIM 的不同之处在于它搜集整合了许多独立的数据库中的数据，包括 OMIM、GenBank、UniGene、UCSC 以及蛋白质序列的慕尼黑信息中心（MIPS）。这样，与 OMIM 相比，GeneCards 对人类疾病的文字描述相对较少，但提供了更多的功能基因组数据。GeneCards 还使用了人类基因组组织（HUGO）的基因符号来确保使用的是标准命名。对于 *MeCP2*，GeneCards 的条目包括染色体上的定位、SwissProt 链接、基于微阵列实验以及 UniGene 统计得到的局部基因表达数据和直系同源基因（图 18-11）。另外一些重要的资源在表 18-8 中列出。一个重要的数据库是人类基因突变数据库（HGMD）（Krawczac 和 Cooper，1997），这个数据库整理了已经公布的对人类遗传疾病有影响的基因变异。

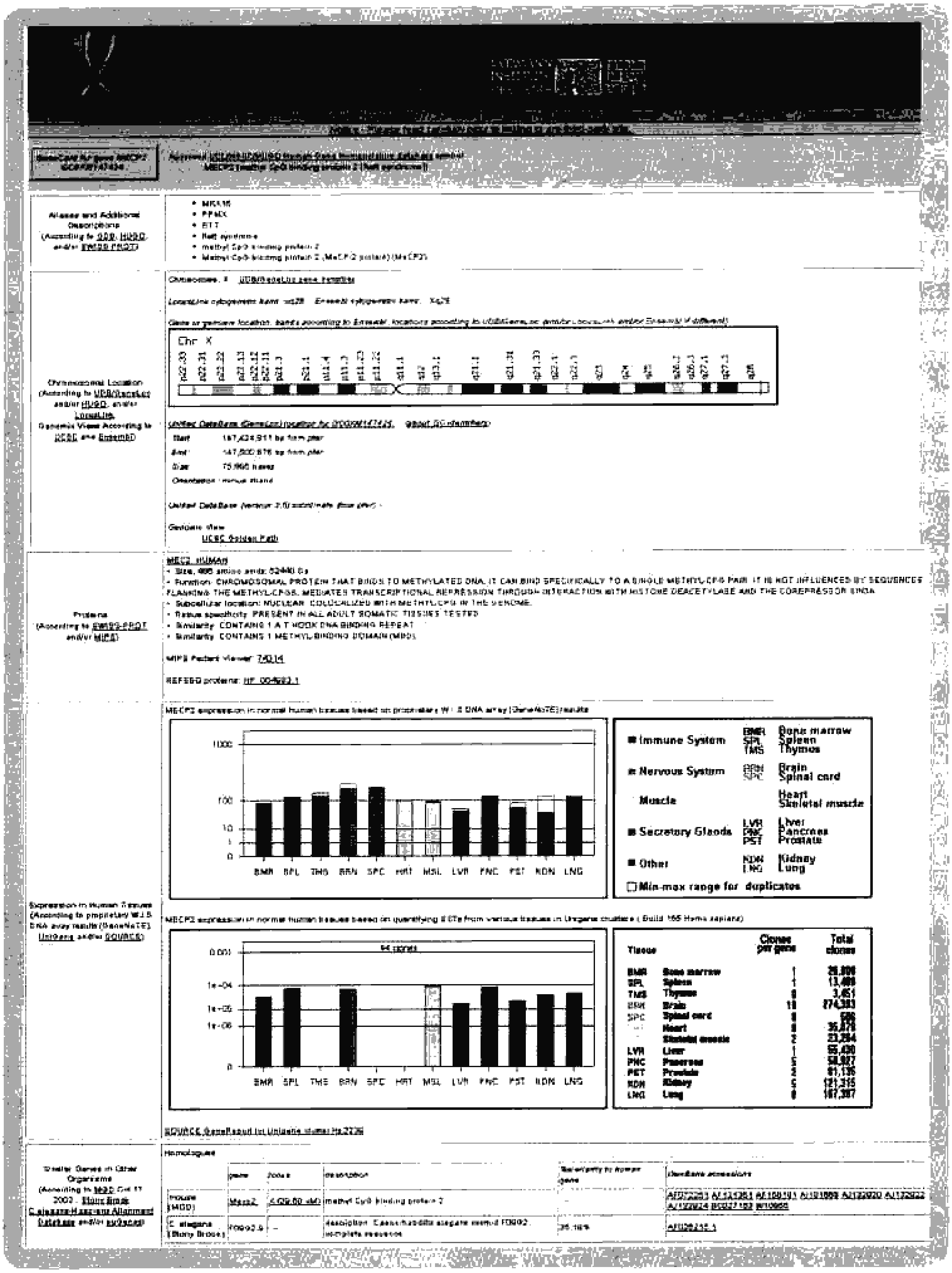


图 18-11 GeneCards 是一个基因组数据库，包含关于基因序列、基因组定位、基因表达、蛋白结构功能以及医学信息的丰富数据。它的站点是 http://bioinfo.weizmann.ac.il/cards/

表 18-8 全球互联网上的重要突变数据库

站点	描述	网址
GeneCards	Weizmann 研究所建立	http://bioinfo.weizmann.ac.il/cards/
HUGO 突变数据库（自主的）	一个包含大量特异性位点突变的数据库	http://ariel.ucs.unimelb.edu.au:80/~cotton/glsdb.htm
人类基因突变数据库	来自 Cardiff 遗传医学研究所	http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html
哺乳动物基因突变数据库（MGMD）序列变异（SVD）	已经公开的由诱变剂诱导的哺乳动物组织突变数据库	http://lisntweb.swan.ac.uk/cmgt/index.htm
数据库计划	在 EMBL-EBI	http://www2.ebi.ac.uk/mutations/

18.6.3 突变数据库

有两类基本的人类疾病基因数据库：全局的数据库和位点特异的数据库。全局的数据库如 OMIM、GeneCards 以及 HMG 试图全面描述所有的疾病相关基因。相反，位点特异数据库（也称突变数据库）更彻底地描述单个基因（有时是几个基因）的变异。这些数据库提供关于单个特定基因遗传方面详尽的专门知识。而且位点特异数据库中的已知突变包含的内容远比全局的数据库深入（Scriver 等，1999）。这样，两类数据库以互补的形式提供服务。

位点特异数据库是一个等位基因变异知识库。现有几百个这样的数据库。一个位点特异数据库的基本要素应该包含如下内容（Scriver 等，1999，2000；Claustres 等，2002）：

① 每个等位基因的惟一标识；

② 数据源的信息；

③ 等位基因的上下文关系；

④ 等位基因的信息（如它的名称、类型以及核苷酸变异）。

突变数据库在收集突变信息中扮演着重要角色，但直到现在对它们的建立还没有一个统一的标准。Claustres 等（2002）调查了包含 262 个位点特异数据库的 94 个网站。他们发现数据的采集、表示、链接、命名以及更新方式存在很大的差异。Scriver 等（1999，2000）描述了突变数据库的目录、结构以及部署的方针。

> 在突变数据库中，一个突变定义为一个等位基因变体（Scriver 等，1999）。等位基因（或者惟一的序列改变）可能是致病的；这样的等位基因倾向于低的发生率。等位基因可能是中性的，在表型上没有任何效应。

① 目前等位基因命名的统一性在增加（Antonarakis，1998；den Dunnen 和 Antonarakis，2000）。例如，起始子蛋氨酸密码子 ATG 中的 A 标记为核苷酸+1。许多这样的规则已经开始明确允许以统一的方式描述突变。

② 对伦理性问题也做了描述，如保守信息机密的义务（Knoppers 和 Laberge，2000）。

③ 提供了构建以及分析位点特异数据库的通用软件（Beroud 等，2000；Brown 和 McKie，2000）。

几个网站提供了访问位点特异数据库的入口（表 18-9）。主要的入口是人类基因组变异协会（HGVS）（图 18-12）。通过它能访问几百个位点特异突变数据库。

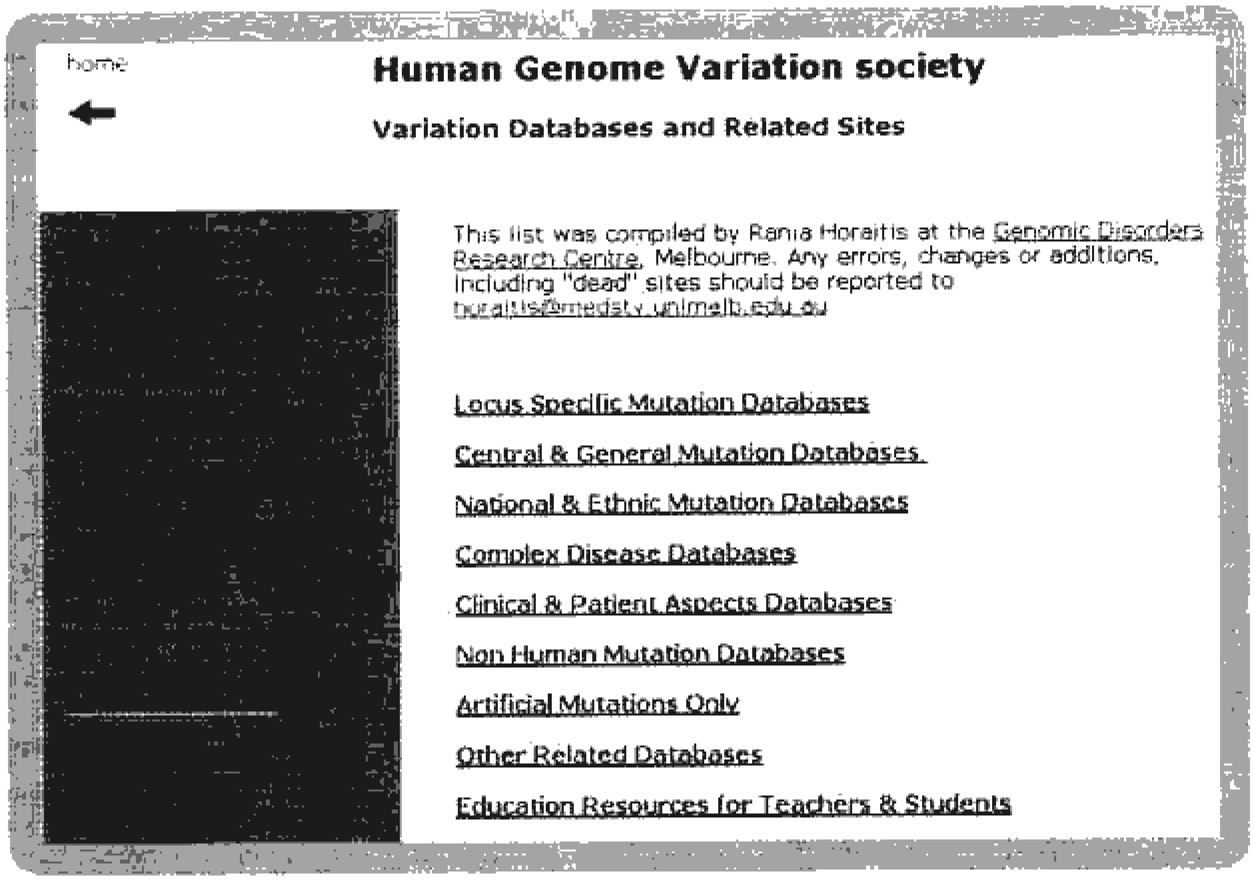


图 18-12 在“人类基因组变异协会”的网页上能链接到许多特异性人类基因突变数据库（http://ariel. ucs. unimelb. edu. au：80/～cotton/dblist. htm）

表 18-9 特异位点数据库的门户网站

站点	描述	网址
GeneDis	来自 Tel Aviv 大学；能对疾病数据库进行比对	http://life2. tau. ac. il/GeneDis/
HUGO 突变数据库（自主的）	一个包含大量特异性位点突变的数据库	http://ariel. ucs. unimelb. edu. au:80/～cotton/glsdb. htm
人类基因突变数据库	来自 Cardiff 遗传医学研究所	http://archive. uwcm. ac. uk/uwcm/mg/hgmd0. html
通用突变数据库	人类基因突变相关的数据库和软件，来自 INSERM	http://www. umd. necker. fr
哺乳动物基因突变数据库（MGMD）	已经公开的由诱变剂诱导的哺乳动物组织突变数据库	http://lisntweb. swan. ac. uk/cmgt/index. htm
序列变异（SVD）数据库计划	在 EMBL-EBI	http://www2. ebi. ac. uk/mutations/
特异基因突变数据库	Engelhardt 分子生物学研究所（俄罗斯）	http://wgen. eimb. relarn. ru/databases/genespec. htm

作为位点特异数据库的一个例子，我们可以调查 RettBASE 数据库。RettBASE 提供了 *MeCP2* 突变相关的临床表征的详细数据（图 18-13）。

18.6.4 单核苷酸多态性（SNPs）与疾病

SNPs 的密度为大约平均每 800bp 有一个。SNPs 可以发生在基因的编码区（导致同义或者不同义的改变）、内含子区、基因调控区，或基因间区域。

Entry ID	Short Citation	Nucleotide change	Amino acid change	Type of sequence change	Domain change location	Phenotype	Mutation/polymorphism	Sex
1271	Klauck, ... Pubmed: 21881244	c.419C>T	p.A140V	Missense	MBD	Not Rett synd. - Unaffected family member		F
1272	Klauck, ... Pubmed: 21881244	c.419C>T	p.A140V	Missense	MBD	Not Rett synd. - X-linked mental retardation, manic-depressive illness		M
1361	Carson, ...	c.422_423insA	p.Y141fs	Frameshift insertion or deletion	MBD	Not known	Mutation associated with disease	F
330	De Bona, ... Pubmed: 10854091	c.423C>G	p.Y141X	Nonsense	MBD	Rett syndrome - Classical	Mutation associated with disease	F
882	Directly submitted	c.423C>G	p.Y141X	Nonsense	MBD	Rett syndrome - Classical	Mutation associated with disease	U
883	Directly submitted	c.423C>G	p.Y141X	Nonsense	MBD	Rett syndrome - Classical	Mutation associated with disease	U
1063	Buyse, ... Pubmed: 20530225	c.423C>G	p.Y141X	Nonsense	MBD	Rett syndrome - Not certain	Mutation associated with disease	F
1134	Watson, ... Pubmed: 21180436	c.423C>G	p.Y141X	Nonsense	MBD	Not Rett synd. - Angelman syndrome	Mutation associated with disease	F
1434	Schwartzman, ... Pubmed: 21412289	c.423C>G	p.Y141X	Nonsense	MBD	Rett syndrome - Male variant	Mutation associated with disease	M
1461	Vacca, ... Pubmed: 21166020	c.423C>G	p.Y141X	Nonsense	MBD	Rett syndrome - Classical	Mutation associated with disease	F
1521	Yamada, ... Pubmed: 21415588	c.423C>G	p.Y141X	Nonsense	MBD	Rett syndrome - classical	Mutation associated with disease	U
45	Hoffbuhr, ... Pubmed: 11402105	c.426C>T	p.F142F	Silent	MBD	Rett syndrome - Not certain	Silent polymorphism	F

图 18-13 RettBASE（http://mecp2.chw.edu.au/mecp2/）是位点特异数据库的一个例子。它描述了已经报道的 *MeCP2* 基因不同位点突变的核酸变异。注意到蛋白的甲基结合结构域（*MBD*）的 423 位的碱基 C 替换成 G 是一个无义突变（如引入一个终止子）。不少病人有这类突变，但是经典的临床表型只包括女性蕾特症患者，男性则是另外一种临床表型

一个个体的 SNPs 基因型有可能提供洞察他或者她感染疾病的可能性。这是因为序列的变异可能和遗传表型相联系。SNP 数据可以通过四种主要的途径提供与疾病相关的信息。

① 功能变异。一个 SNP 可能和编码区的一个不同义替代相关。在这种条件下，蛋白质的氨基酸序列发生改变。据估计人类基因组中有 530 万个 SNPs，每一个的发生频率至少是 10%（Kruglyak 和 Nickerson，2001）。根据这个估计，SNPs 的发生频率大约为每 600bp 有 1 个（对于一个发生频率为 1%的最小等位基因，估计有 1100 万个 SNPs，发生频率大约为每 290bp 对出现 1 个。）

② 调控变异。一个 SNP 可以发生在非编码区域，但是影响基因表达的特性（如时程、区域模式或者生理响应）（Cowles 等，2002）。

③ 联合。SNPs 可以用来进行整个基因组联合的研究（Kruglyak，1999）。在连锁研究中，遗传标记可以在同一家族中用来搜索共同继承的染色体区域。相反，在联合的研究中，

通过研究无关的受影响个体和控制条件来发现基因变异频率上的差异。SNPs表现出的许多类型的序列差异可以用来进行联合研究（Sherry等，2000）。这样，即使SNPs本身没有引起诱导疾病的变异，但它们还可能是有害的突变。

④ 单体型图谱。虽然DNA变异的发生遍布整个染色体，但有些变异则发生在很保守的区域（称作单体型块）。单体型是人群中出现的一个等位基因的组合；更通俗地说，单体型定义为DNA序列变异的模式。任何给定染色体上邻近的SNPs组是以集合或者单体型块的模式来遗传的。这些块包含或多或少的SNPs。一个块的大小也可能达到几百万个碱基对。SNPs可以作为标签来定义整个单体型块的行为。人们正在努力定义人类和其他基因组中单体型块的结构（Patil等，2001；Gabriel等，2002；Phillips等，2003）。这样的单体型图谱（HapMap）可以用来选择在解释不同人种之间以及其他人群之间多样性时能提供丰富信息的SNPs。HapMap计划还可能指明哪些SNPs在检查与疾病表征关联的人类基因组时是最有用的。

SNPs的发现有下面2个来源：

① EST数据库可以用来挖掘单核苷酸差异（Buetow等，1999）；

② 人类基因组序列可以用来分析鉴别SNPs（Taillon-Miller等，1998）。

不管SNPs是怎么被识别的，确定它们是真实的序列差异还是仅仅是人为的原因（如长序列阅读时的序列错误或者聚合酶链反应中人工合成的）造成的是很重要的。

有几个主要的SNP数据库（表18-10）。NCBI上的dbSNP是一个有超过3百万个人SNPs的知识库（Sherry等，2000）。数据库还包括微卫星和小的插入/删除多态性数据。dbSNP现有的内容列在表18-11中。

表18-10 单核苷酸多态性（SNP）资源

资源	注释	网址
dbSNP	在NCBI上	http://www.ncbi.nlm.nih.gov/SNP
人类SNP数据库	在Whitehead研究所	http://www-genome.wi.mit.edu/snp/human
SNP联盟	工业和学术实验室的合作	http://snp.cshl.org

表18-11 dbSNP数据统计（NCBI，2002年11月）

功能分类	SNP数目	基因数目
基因区	291459	26482
等位同义重叠群	12322	7147
等位不同义重叠群	16251	8496
不翻译区	131987	13208
内含子	904573	22113
剪切位点	277	268
等位全同重叠群	28491	11621
编码：意义不明	13501	3584

我们可以在NCBI的dbSNP中搜索*MeCP2*变体。输出结果包含来自于mRNA比对和重叠群注释的多态性相关数据（图18-14）。SNPs在图和表格中以彩色标记用来表示编码区的变异（同义的或者不同义的）以及不翻译区、内含子、剪切位点的变化。

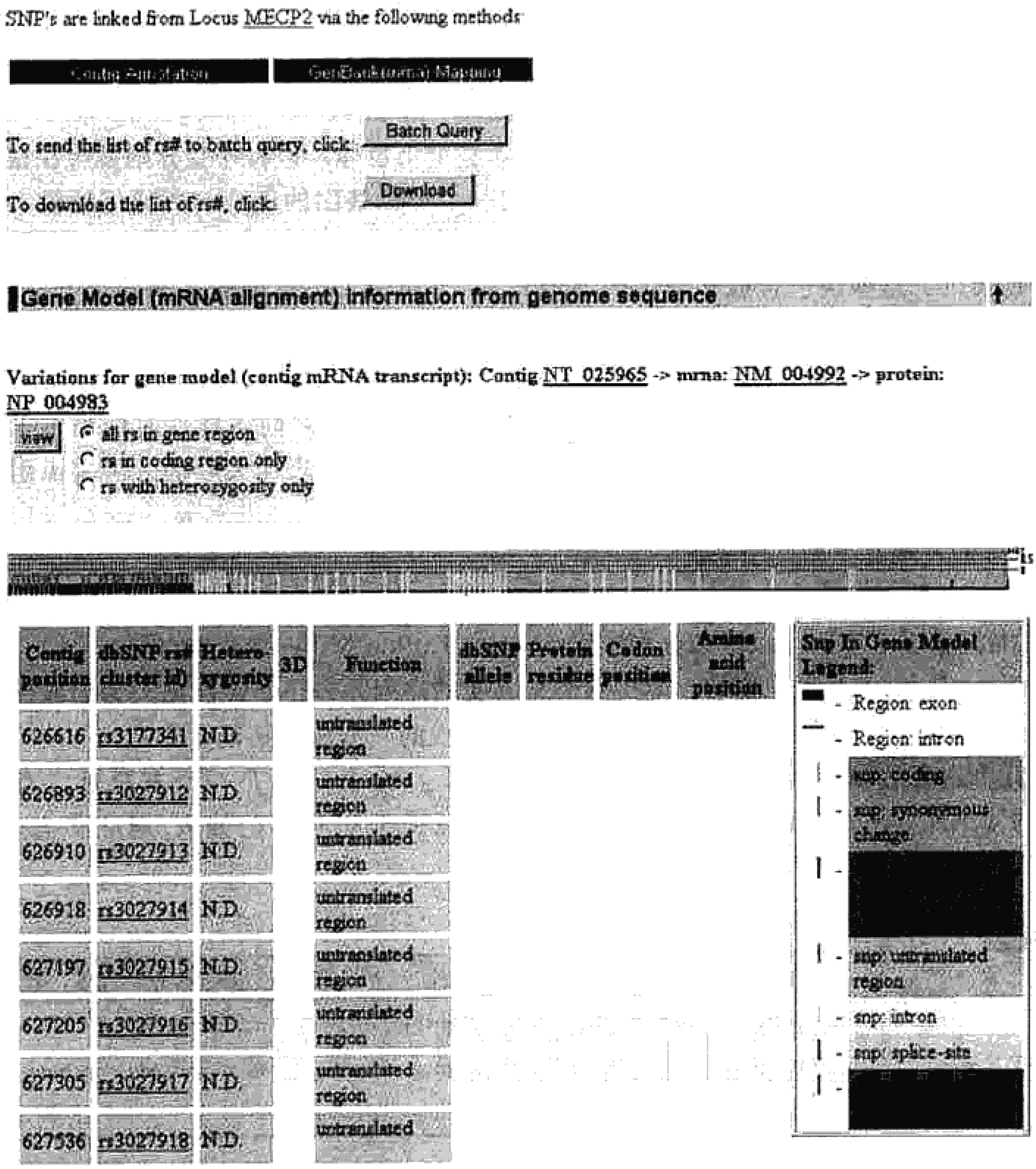


图 18-14 NCBI 上的单核苷酸多态性数据库（dbSNP）显示出 SNPs 和 *MeCP2* 基因的联系

18.7 复杂疾病

与单基因疾病相比，复杂疾病在人群中更为常见（Todd，2001）。这个特点并不背离简单、离散的孟德尔方式。复杂疾病的例子有哮喘、孤独症（框 18-3）、糖尿病、高血压、肥胖症，以及骨质疏松症。很可能人类的多数疾病都和多个基因相关。复杂疾病表现出以下特性。

① 涉及多个基因。多个基因突变的组合是这种疾病的主要因素。在单基因疾病中，即使是位点的修饰，一个基因对疾病的表征也会有巨大的影响。

② 复杂疾病涉及多个基因的组合效果，但是它们也可由那些提高患病可能性的环境因素和行为所引起。

③ 复杂疾病是非孟德尔类的：它们表现出家族凝聚而不是分离。例如，孤独症是一种高

度可遗传疾病（如果同卵双生双胞胎中的一个患病，那么另外一个也会有很高的患病概率）。

④ 易感性等位基因具有较高的种群频率；也就是说，复杂疾病通常比单基因疾病更加常见。镰刀型红细胞贫血症作为一种单基因疾病在美籍非洲人群中的发生率异乎寻常的高，但是这种杂合状态赋予了一种自然选择上的优势（见框 18-1）。

⑤ 易感性等位基因的外显率低。外显率是显性或者纯合的隐性基因在一定人群中表现出它的特征表型的频率。在极端情形下，这是一种全或无的现象，即基因型表现或者不表现。在复杂疾病中，部分的外显性是普遍的。

框 18-3 孤独症：未知病因的复杂疾病

孤独症（OMIM' 209850）是一种三岁前就发作的终身神经疾病（Kanner，1943；Rapin 和 Katzman，1998；Bailey 等，1996；Ciaranello，1995；Piven，1997；Rapin，1997；Waterhouse 等，1996；Gillberg 和 Coleman，2000；Lord 等，2000）。它有三个特征缺陷：①患病个体失去正常相互社交的能力；②语言和交流能力减弱；③有限而固定的兴趣和活动。孤独症儿童的行为从幼年起就开始不正常，明显缺乏想像力。大约 30% 的孤独症儿童看起来发育正常，但随后在 18～24 个月的年龄时会经历一个时期的语言技能退化。大约 10% 的孤独症患者在数学计算、机械式记忆，或者音乐表演方面具有专家般的超常能力。孤独症常伴随着癫痫；到成人期为止，大约 1/3 的孤独症患者至少有两种无缘无故的癫痫（Rossi 等，1995；Olsson 等，1988；Volkmar 和 Nelson，1990）。

据估计孤独症的流行程度大约为每 100 人中有 0.2～2 人（Rapin 和 Katzman，1998；Fombonne，1999；Smalley 等，1988；Gillberg 和 Wing 等，1999）。男性的患病率大约是女性的 3～4 倍。

对孤独症的病因还不清楚，但是强有力的证据表明这种疾病是遗传的（Smalley 等，1988；Szatmari 等，1998；Turner 等，2000）。同卵双生子之间的一致性大约是 60%，如果共同患病的双生子确定为典型的孤独症或者更一般的在社交、语言以及认识能力上的损害，那么这种一致性就大于 90%（Bailey 等，1995）。孤独症比其他大多数常见的神经性疾病（如精神分裂症和忧郁症）有更强大的遗传基础。尽管有迹象表明与孤独症相关的基因在 15 号染色体的长臂上（Cook 等，1997；Bundey 等，1994；Flejter 等，1996；Baker 等，1994），但是多数遗传连锁研究没有发现有显著关联的基因。几种用微卫星标记的方法在基因组范围内搜索也没有找到与某条染色体的显著连锁，只是发现了一些潜在易感性区域（国际孤独症分子遗传研究协会，1998；Phlippe 等，1999；Lamb 等，2000）。这些研究的主要结论是孤独症很有可能是由多个相互作用的基因变异引起（Gillberg，1998）。已知的医学治疗条件会影响中枢神经系统，像 X 染色体易碎症和癫痫，它们在孤独症病例中占 10%～30%（Barton 和 Volmar，1998 年）。

我们怎样确定复杂疾病的原因呢？有很多方法用来寻找治疗手段（或者最终的治愈手段），我们将以孤独症为明确的例子（框 18-3）考虑其中的几种。

① 对很多复杂疾病来说异质性是一个问题。在孤独症的情形下，临床行为学研究确定了这种疾病的表征。一些病人有很严重的智力障碍而另一些则具有正常的智商。语言技能延迟或缺乏的程度有很大的可变性。这样，“孤独症”这个术语就有可能归属到一组表征上相关的、诊断界限有重叠的疾病中。

② 对风险因素的研究是基于对流行病学的研究。

③ 孤独症的病因或许可以通过遗传学研究找寻，例如这种疾病的流行性在同卵和异卵双生子中的流行病学检测。孪生子研究和家族研究显示出孤独症具有很强的遗传成分。

④ 遗传缺陷可以通过对染色体分析找寻。实际上，一些染色体易位的病人有类似于孤独症的征象。

⑤ 连锁和联合分析可能被使用。多家族连锁研究表明第 7 条和第 15 条染色体以及其他位点与孤独症的病因相关。图 18-15 是几个基因组连锁研究（Lamb 等，2000）中暗示的第

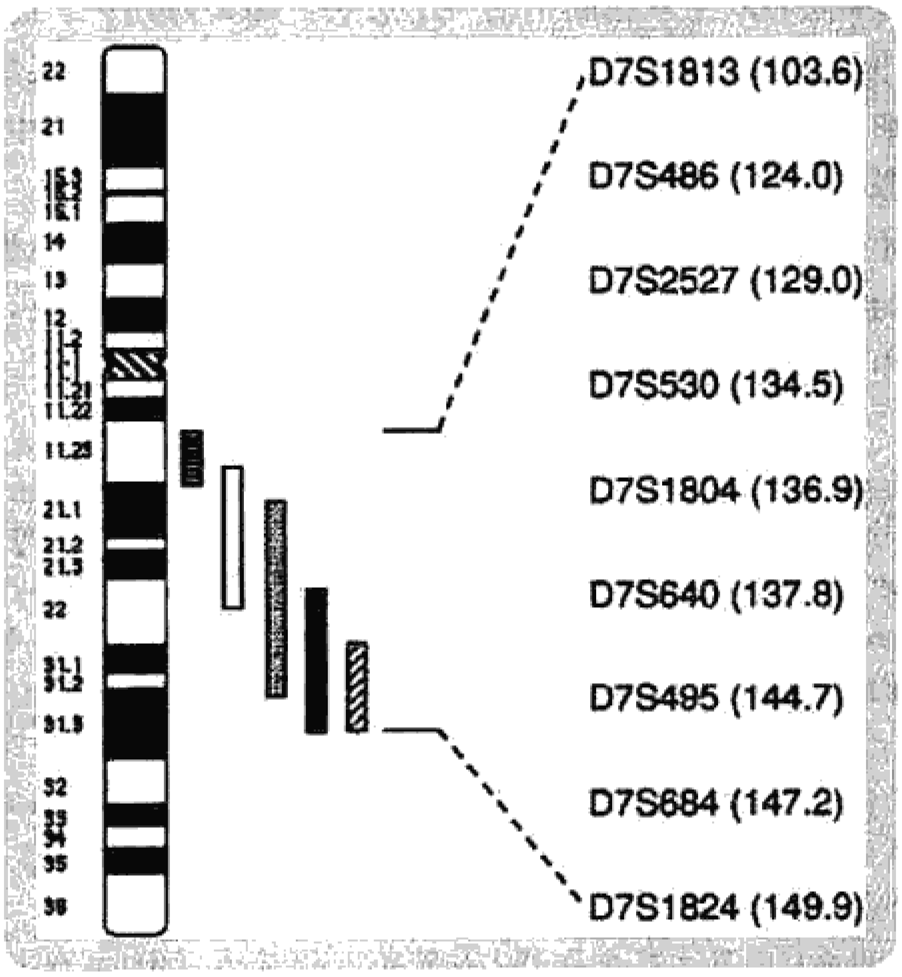


图 18-15 第7条染色体上基于连锁研究发现的孤独症病因学区域的示意图。像NCBI图浏览器这样的生物信息学工具对于确定在特定染色体区域的基因很有用。来自Lamb等（2000），已授权使用

7条染色体上疾病相关基因区域的示意图。该示意图和NCBI的图查看器（图17-2）的形式相似并且包含Marshfield图中所显示的从染色体短臂到端粒的标记距离（以cm为单位）。

⑥ 药理学方法可以用来发现改善疾病临床症状的药物。神经病理学研究还可以确定脑部的患病区域。

⑦ 可以建立动物模型，特别是在致病基因被确定以后。

对于复杂疾病的一个重要研究方法是发现突变时能导致疾病的基因。一个相关的方法就是确定体内致病的突变基因的序列。例如，用微阵列或相关技术获得的基因表达谱可以用来发现影响组织并导致疾病的转录反应。这可能会对疾病的药理学干涉或是分子标记识别提供途径（Colantuoni等，2000；Purcell等，2001）。相同地，蛋白质学方法也可以用来治疗疾病。

18.8 疾病中染色体异常性分析

人体内大范围的染色体异常通常会导致疾病的发生。这些大范围的异常包括染色体的非整倍体性，例如三染色体性和单染色体性。13号染色体三体性［帕韬综合征（patau syndrome)］、18号染色体三体性［爱德华综合征（Edwards syndrome)］、21号染色体三体性［先天愚型（down syndrome)］是常染色体中出现的染色体三体性（表18-12），这种染色体的三体性可以使生命存活（不过，13号染色体三体性和18号染色体三体性的婴儿出生后第一年多数会死亡）。染色体异常导致许多的疾病，例如猫叫综合征（Cri-du-chat syndrome)、Angelman综合征、普-威安吉尔曼综合征（Prader Willi syndrome)、Smith-Magenis综合征和各种由于染色体组中某些区域基因的缺失或者插入导致的精神疾病。某些癌症的发生也可能是由于某些基因的缺失或者插入导致的。Lupski（1998）定义了染色体异常（genomic disorders）的，即染色体结构的改变导致疾病的发生。

表 18-12 存活婴儿染色体出现不正常的概率

异常性	不正常问题	概率
常染色体	13 号染色体三体性	1/15000
	18 号染色体三体性	1/5000
	21 号染色体三体性(先天愚型)	1/600
性染色体	克兰费尔特综合征(47,XXY)	男性 1/700
	XYY 综合征(47,XYY)	男性 1/800
	三倍体 X 综合征(47,XXX)	女性 1/1000
	特纳综合征(45,X 或者 45X/46XX 或者 45X/46,XY 或者异染色体 Xq)	女性 1/1500

注：资料来自 Beaudet 等（2001），授权使用。

染色体异常是正常的人类发育过程中的非常普通的特征。人类有非常低的生育能力，甚至和其他的哺乳动物相比其生育能力也很低，大约 50%～75%流产是因人类的观念影响。人类这么低的生育力也是由于染色体异常所致（Wells 和 Delhanty，2000；Voullaire 等，2000）。

① 已经有了一个孩子的妇女在一个月经期内有 25%的机会可以怀孕。

② 52%的怀孕妇女有一次早产。

③ 在体外受精中，被确认怀孕的妇女受孕后的前两周 30%流产。

④ 发生在受孕后的第 12 周或者更早的自然流产中，大约 60%都是因为染色体的非整倍体性造成的，这也说明早期的受孕失败主要是因为胎儿的染色体发生了致命性的改变。

在胎儿的早期发育中，最主要的染色体失常是获得或者丢失了整条染色体。这种染色体的结构改变可以通过标准的细胞生成方法检测到，该方法包括染色体组型分析（karyotype analysis）和原位杂交荧光染色（FISH）。这些技术也可以用来获得染色体的一些普通现象，例如染色体大范围的复制、丢失或者包含上百万碱基对的重组。

研究染色体在正常发育时表现异常的一个新技术是利用染色体组芯片（框 18-4）。Wells 和 Delhanty（2000）扩增了 12 个正常发育的人胚胎的染色体 DNA，这些人胚胎是通过体外受精 3d 后获得的。他们分析了从这些胚胎中获得的 64 个单细胞的 DNA，通过对每个细胞

框 18-4 染色体组芯片

如今用来检测染色体异常的主要技术为染色体组芯片技术（Hermsen 等，2001；Pollak 等，1999；Pinkel 等，1998；Lichter 等，2000），这种技术现在被称为染色体组杂交比较（comparative genomic hybridization）。在单体时，通过细菌人工染色体（BAC）克隆就可知道染色体的位点，然后把它们放到玻璃载片的表面。每个 BAC 的长度在 200kb 左右，并且 BAC 克隆的空间间隔大约为 0.5～10Mb。这种技术为芯片上的 BAC 克隆提供了增长空间，因为它们在芯片上比在染色体上有更大的空间。

染色体 DNA 是从病人和对照样本中分别纯化出来的，并且染上绿色或者红色荧光（如 Cy3 和 Cy5），然后在芯片上共杂交。最后颜色的比值就反映出染色体 DNA 和每个 BAC 杂交的程度，并且在多数情况下这个比值为 1∶1。因此，对那些染色体自我复制或者染色体片段缺失的情况，就可以通过所得比值的偏差情况检测出来。

染色体组芯片类似于用来做基因表达分析的 DNA 芯片（第 6 章和第 7 章）：在每个例子中，DNA 都被固定在玻璃载片或者其他固体的载片上，并且都需要给对照和实验样本染上红色和绿色的荧光，以及都需要对芯片表面的 DNA 进行共杂交，杂交后需要对芯片进行清洗，并且每个格点的值要用扫描仪来测得。但是，它们之间也有不同，主要就是在染色体组芯片中，样本包含的是各种片段的染色体 DNA（不是 RNA 或者 mRNA），并且从芯片中要获得的信息是要找到整个染色体组 DNA 中的复制或者缺失部分，而不是像 DNA 芯片那样要测得单个基因的表达水平。

中染色体DNA的比较，他们获得了标准对照，从而他们能够测得每个样本中以10Mb作为长度间隔的染色体复制数量的比例。在12个表型正常的胚胎中，有9个胚胎中含有非整倍体染色体组的细胞。这些异常发生在常染色单体中，甚至还有三个胚胎为缺对染色体（染色体中完全丢失了复制的部分）胚胎。Voullaire等（2000）独立地利用同样的方法，扩增了从9个正常胚胎中获得的65个细胞的染色体DNA。在9个胚胎中，仅有3个胚胎自始至终具有正常的染色体谱。他们从其他的胚胎中观察出了染色体的非整倍性，其中包括单倍体、三倍体、染色体的镶嵌和破损。在这两个例子中，报道的"无序"胚胎例子中每个被检测出含有大量非整倍体染色体的细胞都是因为染色体的随机隔离。

综上所述，这些研究表明正常受精后的胚胎多数都含有非整倍性染色体，并且几乎所有的这些染色体异常都能导致胚胎的早期死亡。这些染色体失常主要是因为体外受精的成功率比较低。生物信息学工具（如染色体组芯片）可以完成在保证染色体完整的情况下，对胚胎发育过程中有显著复杂性的基因组宽度（genome-wide）的观察。这种技术还可以用来分析人或其他物种（包括那些相对高或者低生育能力的）染色体改变后的功能改变。

人类疾病系统水平的生物信息学资源（细胞器和通路数据库）真核细胞可以进一步分为各种细胞器，例如细胞核、内质网、高尔基体、过氧化酶体和线粒体等。每个细胞器执行各自特殊的功能，它们都含有细胞生存所需要必需反应的蛋白酶，拥有各自的代谢进程，并且隔离有害物质。我们已经从基因和基因产物的层面来考虑人类的疾病，下一步需要做的就是检查疾病和具有较高组织水平的细胞器或者通路的关系。一系列的网络资源见表18-13。

表18-13 在系统水平上研究人类疾病的一些网络资源

资源	评论	网址
隐性遗传疾病数据库	人类隐性遗传疾病数据库	http://ken.mitton.com/ern/lensbase.html
国际视网膜数据库	四个数据库：蛋白质数据库、突变数据库、疾病数据库和动物模型数据库	http://www.retina-international.org/sci-news/database.htm
国际心血管疾病数据库	关于危险因子、死亡速率的数据和文献	http://www.aihw.gov.au/cvdhtml/cvd-menu.htm
肠胃疾病数据库	数十种关于肠胃疾病的资源	http://www.gastro.net.au/gastrodiseases/
肌体营养失调数据库	关于基因、疾病和突变信息的数据库和资源	http://www.dmd.nl/
线粒体图数据库	人类线粒体染色体组数据库	http://www.mitomap.org/
过氧化酶资源	医生、科学家和其他一些人总结的数据库	http://www.peroxisome.org/

首先，我们来讨论线粒体。这个细胞器在20世纪40年代被发现，主要与呼吸作用有关，并且线粒体中含有DNA（由Nass等在1963第一次报道）。但是，直到1988年，由于线粒体DNA的突变造成疾病才首次报道出来（Wallace等，1988；Holt等，1988）。现在，已经有100多种关于线粒体内点突变造成的人类疾病被报道出来（见DiMauro和Schon的综述，2001；Schon，2001）。线粒体染色体编码37个基因，其中的任何一个都和人类疾病有关系（见图12-7和图12-8）。图18-16显示了线粒体染色体的发病图谱。

> 多数线粒体中的蛋白都是细胞核中的基因产物，并且多数线粒体疾病都是因为细胞核基因的突变造成的。

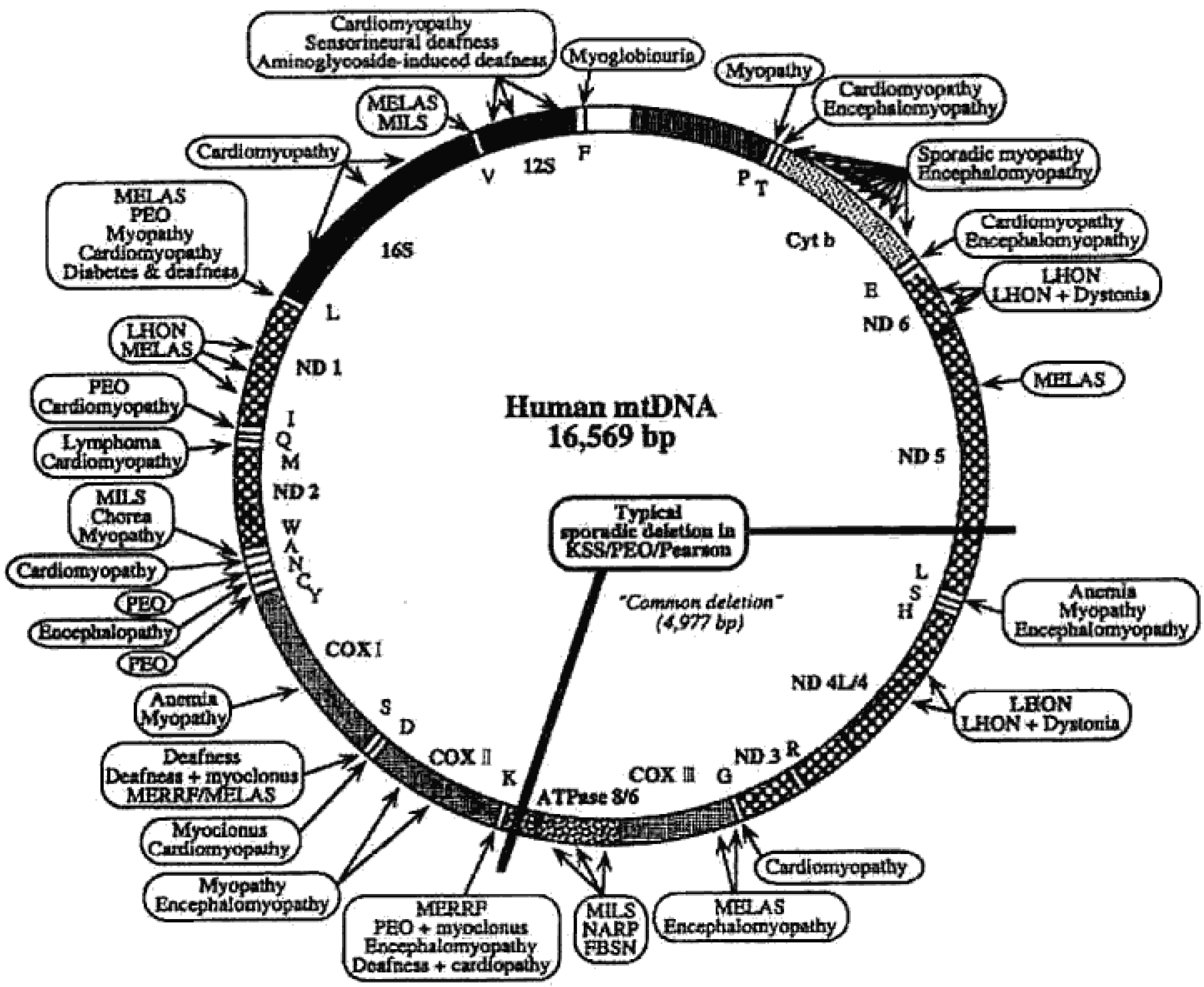


图 18-16 人类线粒体染色体的发病图谱。ND—编码 7 个复合体 I 亚单位的基因；COX—3 个 cytochrome 氧化物复合体；Cyt b—Cytochrome b；ATP 酶 6 和 ATP 酶 8—2 个 ATP 合成酶亚单位；12S 和 16S—核糖体 RNAs（22 个转运 RNAs 通过对应氨基酸的单字母被鉴别出来）；FBSN—家族的有双边条纹的坏疽；KSS—Kearns-Sayre 综合征；LHON—Leber 遗传精神病；MELAS—线粒体脑肌症、乳酸毒症；MERRF—肌阵挛癫痫症和纤维粗糙发红症；MILS—母本遗传异体综合征；NARP—视网膜色点、运动失调神经病患者；PEO—外在眼肌瘫痪。来自 DiMauro 和 Schon（2001），授权使用

线粒体遗传特性与孟德尔遗传法则主要有三个方面的不同（DiMauro 和 Schon，2001）。

> 通常所有的线粒体染色体都相同，这种情况被称为同形。致命的突变可能是异质体的（正常染色体和突变染色体的混合体）。

① 线粒体 DNA 是母性遗传的。在胚胎中，线粒体主要来自卵子，同时精子的线粒体在进入受精卵前就被分解掉。因此，如果母亲含有突变的线粒体 DNA，那么就会遗传给她的子女，不过只有她的女儿才能把这种突变传递到下一代。

② 细胞核基因存在两个等位点（一个来自母本一个来自父本），而线粒体基因在一个细胞中存在成百上千个拷贝（一个典型的线粒体会包含线粒体染色体的 10 个拷贝）。任何一个拷贝都可能有多种正常和突变线粒体染色体组的比值。而我们需要的是判断导致某些疾病是否会发生的线粒体染色体突变的临界极限。

③ 在细胞分裂时，线粒体中染色体突变的部分可以改变，这样就会影响线粒体异常的

表型的表达。临床表明，线粒体异常可以发生在任何时间和机体内的任何部位。现在已知有很大一部分疾病都和线粒体 DNA 的突变有关。

MITOMAP 是一个全面的关于线粒体染色体的数据库（见表 18-13 的网址）。在该数据库中列出了各种各样的关于线粒体染色体组的突变和多态现象，包含了所有已知的遗传机制（倒置、插入、删除等）。

18.9 在模式生物中研究和寻找人类疾病基因

在动物中研究人类疾病的一个重要的网络资源是孟德尔遗传在线（OMIA），网址为 http://www.angis.org.au/Databases/BIRX/omia/。

我们主要是通过在其他物种中研究人类疾病相关的基因和基因产物，从而了解人类疾病的病理生理学。我们已知基因突变会导致疾病发生，所以研究思路主要是先在人类细胞中找到具有某种功能的异常蛋白，然后在其他模式生物中突变翻译这些蛋白的基因，从而使模式生物患病。一旦在模式生物中找到和人类疾病有关的基因，把这些基因从模式生物中敲除或者进行其他的处理，就可以得到模式生物某些特殊突变的表型。

在果蝇的序列刚被测出来时，大约 65% 的人类疾病基因被证明和果蝇的基因具有同源性（Ahringer，1997）。

在模式生物中研究人类疾病基因有一个基本问题，即如何鉴别已知的人类疾病基因在模式生物中是否有同源基因。这种技术即使对鉴别那些祖先不同的生物物种的突变也是有用的。由 55 人组成的小组合作对果蝇（*Drosophila melanogaster*）、线虫（*Caenorhabditis elegans*）和酵母（*Saccharomysces cerevisiae*）基因组序列进行分析（Rubin 等，2001），他们通过突变、替换、扩增和缺失的方法鉴别出 289 个和人类疾病有关的基因。在这些基因中，177 个（61%）基因和果蝇基因有同源性。这些数据显示在图 18-17 中，图中显示出在果蝇、线虫和酵母中和人类基因中有同源性的是按癌症、神经综合征、心血管疾病、内分泌疾病和其他疾病类型来分类的。Reiter 等（2001）的研究成果是把这些模式生物中与人类疾病有关的基因扩展到 929 个，其中 714 个基因的产物和 548 个果蝇蛋白序列匹配（表 18-14）。Reiter 等（2001）的数据存放在 Homophlia（一个关于现代人/果蝇疾病的数据库）上。

表 18-14 基于人类疾病表型的 714 个果蝇基因的聚类

异　常	基因数目	异　常	基因数目
神经学方面		免疫学方面	
神经肌肉方面	20	间接补体	11
神经精神方面	9	其他方面	22
CNS/发育方面	8	总计	33
CNS/共济失调	9	血液方面	
弱智	6	红细胞方面	29
其他方面	22	卟啉症	7
总计	74	血小板方面	6
		总计	42

续表

异常	基因数目
内分泌方面	
糖尿病	10
其他方面	40
总计	50
耳聋方面	
综合征方面	7
非综合征方面	6
总计	13
心血管方面	
心肌症	10
控制缺失	4
高血压	7
动脉硬化症	3
血管畸形	2
总计	26
眼科方面	
外在部分	
无虹膜	1
Rieger 综合征	1
间叶细胞发育失调	2
虹膜发育失调	2
角膜营养失调	2
白内障	3
青光眼	2
小计	13
视网膜部分	
视网膜营养失调	1
无脉络膜	1
色盲	4
锥形体营养失调	2
锥形体杆营养失调	1
夜盲症	8
Leber 先天黑内障	2
斑点营养失调	4
色素性视网膜炎	7
小计	30
总计	43
肺方面	4
肠胃方面	13
肾方面	13

异常	基因数目
异常凝结方面	28
对身体有害基因	
脑方面	3
胸方面	4
结肠方面	11
其他肠胃问题	3
泌尿生殖器方面	5
gynocological	3
内分泌方面	3
皮肤病方面	3
色素性燥皮症	6
其他/恶性毒瘤	9
血液损害	29
总计	79
骨骼发育方面	
颅骨结合	5
骨骼发育异常	13
其他方面	8
总计	26
软组织方面	2
结缔组织方面	18
皮肤病学方面	25
代谢/线粒体	123
药理学方面	12
过氧物酶体	9
储存方面	
糖原质储存	11
脂类储存	13
黏多糖	10
其他方面	3
总计	37
多向性发育方面	
生长、免疫、癌症	7
细胞程序性死亡	1
其他方面	27
总计	35
其他复合物方面	9
最后总计	714

注：资料来自 Reiter 等（2001），授权使用。

在模式生物中对人类疾病相关基因编制目录是非常重要的，我们可以根据目录为这些基因创建功能分析。除了在果蝇、酵母和线虫中获得了结果，在对一些真核生物，如裂殖酵母（*Schizosaccharomyces pombe*）（Wood 等，2001）和拟南芥（*Arabidopsis*）的研究中也有了重大的突破。对裂殖酵母的分析，从它的基因组中鉴别出来和人类疾病基因同源

Homophlia 的网址为 http://homophila.sdsc.edu/。

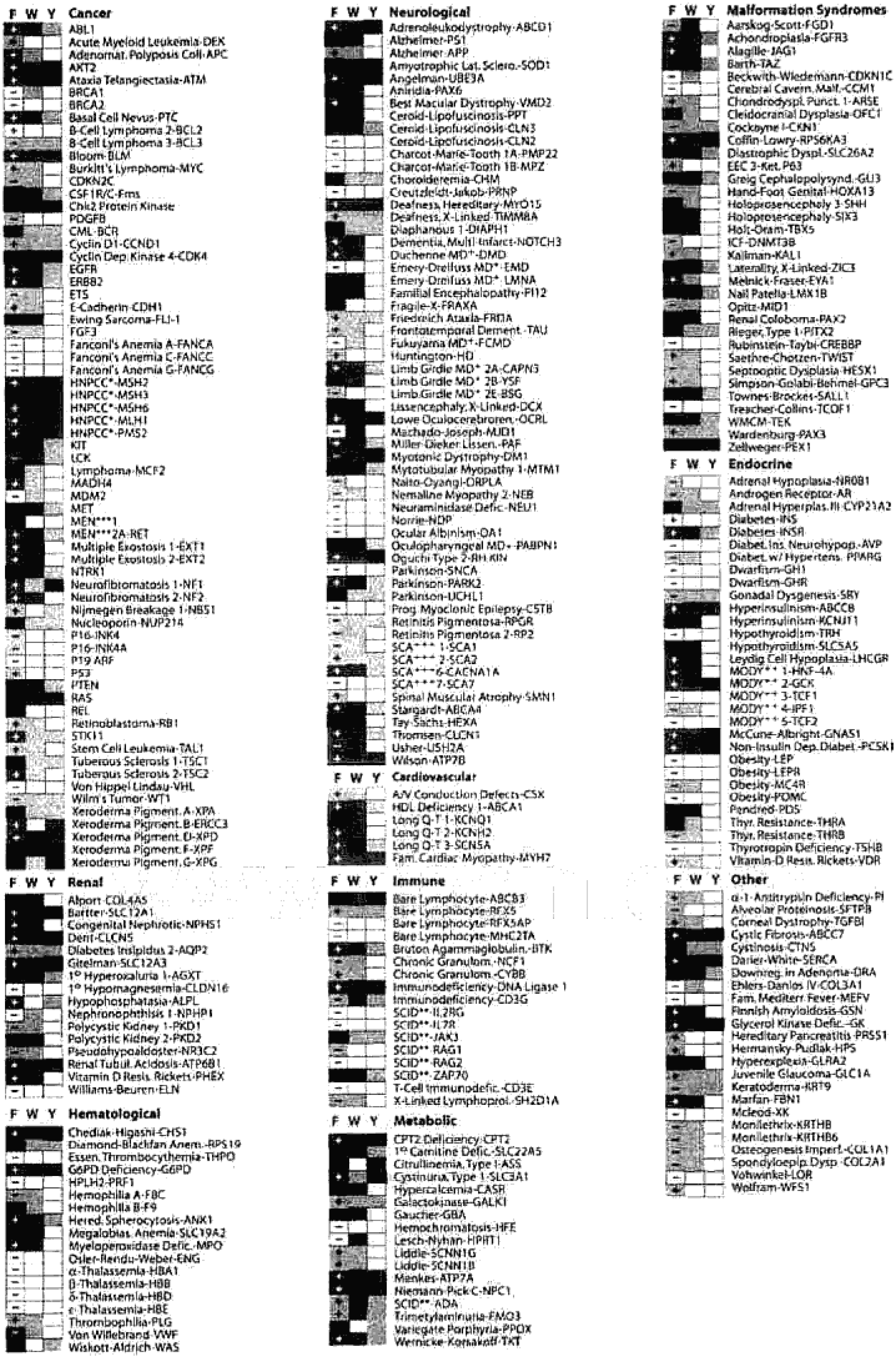


图 18-17 将 289 个人类疾病基因编码的蛋白和果蝇（F）、酵母（Y）和线虫（W）的整个基因组 38860 个蛋白进行序列比对，所用工具为 blastp。匹配的数据是根据比对统计的重要性水平列出。白框表示 E 值大于 1×10^{-6}（没有相似度或相似度较低）。亮灰色的框表示 E 值从 1×10^{-6} 到 1×10^{-40}，红框表示 E 值从 1×10^{-40} 到 1×10^{-100}。黑灰色的框表示 E 值小于 1×10^{-100}。“+”表示果蝇蛋白和人类蛋白功能相同（基于的标准包括序列相似、InterPro 结构域组成等生物学证据）。“—”表示没有证据证明该蛋白和人类蛋白功能相同。引自 Rubin 等（2001），使用得到授权

的基因，这些基因涉及到癌症基因（表 18-15）、各种精神疾病、代谢混乱以及其他一些异常有关的基因（表 18-16）。

表 18-15 与人类癌症基因相关的裂殖酵母基因

人类癌症基因	分数	裂殖酵母基因/产物	体系名称
色素性燥皮症 D；*XPD*	$<1\times10^{-100}$	*rad15*，*rhp3*	SPAC1D4. 12
色素性燥皮症 B；*XPB*	$<1\times10^{-100}$	*rad25*	SPAC17A5. 06
遗传性非息肉性大肠癌(HNPCC)；*MSH2*	$<1\times10^{-100}$	*rad16*，*rad10*，*rad20*，*swi9*	SPBC24C6. 12C
色素性燥皮症 F；*XPF*	$<1\times10^{-100}$	*cdc17*	SPCC970. 01
遗传性非息肉性大肠癌；*PMS2*	$<1\times10^{-100}$	*pms1*	SPAC57A10. 13C
遗传性非息肉病性大肠癌；*MSH6*	$<1\times10^{-100}$	*msh6*	SPAC19G12. 02C
遗传性非息肉病性大肠癌；*MSH3*	$<1\times10^{-100}$	*swi4*	SPCC285. 16C
遗传性非息肉病性大肠癌；*MLH1*	$<1\times10^{-100}$	*mlh1*	SPAC8F11. 03
血液 Chediak-Higashi 综合征；*CHS1*	$<1\times10^{-100}$	—	SPBC1703. 4
Darier-white 疾病；*SERCA*	$<1\times10^{-100}$	*pgak*	SPBC28E12. 06C
Bloom 综合征；*BLM*	$<1\times10^{-100}$	*hus2*，*rqh1*，*rad12*	SPBC31E1. 02C
共济失调微血管扩张症候群；*ATM*	$<1\times10^{-100}$	*Tel1*	SPAC2G11. 12
色素性燥皮症 G；*XPG*	$<1\times10^{-40}$	*Rad13*	SPBC3E7. 08C
结节性硬化症 2；*TSC2*	$<1\times10^{-40}$	—	SPAC630. 13C
免疫裸淋巴细胞；*ABCB3*	$<1\times10^{-40}$	—	SPBC9B6. 09C
腺瘤下调；*DRA*	$<1\times10^{-40}$	—	SPAC869. 05C
黑钻石贫血；*RPS19*	$<1\times10^{-40}$	*rps19*	SPBC649. 02
可凯因症；*CKN1*	$<1\times10^{-40}$	—	SPBC577. 09
RAS	$<1\times10^{-40}$	*Ste5*，*ras1*	SPAC17H9. 09C
细胞周期蛋白酶 4；*CDK4*	$<1\times10^{-40}$	*Cdc2*	SPBC11B10. 09
CDK2 蛋白酶	$<1\times10^{-40}$	*Cds1*	SPCC18B5. 11C
AKT2	$<1\times10^{-40}$	*Pck2*，*sts5*，*pkc1*	SPBC12D12. 04C

注：图中的得分是用 BLAST 工具搜索得到的 *E* 值，表中$<1\times10^{-40}$的指分数在 1×10^{-40} 和 1×10^{-100} 之间。资料来自 Wood 等（2002），授权使用。

我们发现与人类癌症有关的基因同样在真菌中也会出现，例如在真菌中包含编码 DNA 损坏和修复以及控制细胞周期的蛋白。也许我们会惊讶于关于神经错乱的基因同样会出现在单细胞的真菌中。但是，关于这个问题的解释却很简单，因为神经原细胞是特别容易受到影响的细胞类型，它们有惟一的代谢通路。例如，大多数的溶酶体异常都是因为控制溶酶体功能或者溶酶体胞内运输的酶的缺失造成的。多数器官系统在这种情况下会有自我调节过程，但是精神疾病（如智障）就是这种异常情况的基本结果。溶酶体在细胞中主要执行的是催化功能。在真菌中，液泡执行了与溶酶体功能相似的功能，并且许多和人类同源的真菌液泡蛋白已经被鉴别出来了。

对我们来说，从模式生物中鉴别出和人类疾病基因同源的基因的重要作用是什么呢？除去对现在已知的同源基因编排目录，下一步需要做的就是要获得这些人类疾病基因中比较保守的氨基酸残基的突变信息。Miller 和 Kumar（2001）选择了 7 个基因（这 7 个基因的共同特点就是当它们突变后就会导致人类疾病），它们分别为：胆囊纤维症跨膜调控子（CFTR），葡萄糖 6 磷酸脱氢酶（G6PD）基因，神经中枢细胞支持分子 L1（L1CAM），苯基丙氨酸羟化酶（PAH）基因，成对盒 6（PAX6），X-连锁视网膜基因（RS1）和有结节的硬化症基因（TSC2）。对于这些基因中的每个基因，有两种资源可以利用：在病人中的特殊位点突变数据库和各种多细胞动物（如灵长类动物、啮齿类动物、鱼类、昆虫类和线虫类）同源蛋白的序列。他们对这 7 个基因的每个基因做了多序列比对，从而检验 H_0 假设（即在每个基因中点突变的发生是随机的）。通过统计检验（χ^2 分析），他们得到多数可以导致人类

疾病的氨基酸突变是保守的（至少在哺乳动物中）。

表 18-16 与人类疾病基因相关的裂殖酵母基因

人类疾病及疾病基因	疾病类型	分数	裂殖酵母基因/产物
威尔逊病；*ATP7B*	新陈代谢方面	$<1\times10^{-100}$	P 型铜 ATP 酶
非胰岛素依赖性糖尿病；*PCSK1*	新陈代谢方面	$<1\times10^{-100}$	Krp1，驱动蛋白相关
胰岛功能亢进；*ABCC8*	新陈代谢方面	$<1\times10^{-100}$	ABC 转运子
G6PD 缺乏；G6PD	新陈代谢方面	$<1\times10^{-100}$	Zwf1 GP6 脱氢酶
瓜氨酸血Ⅰ型；*ASS*	新陈代谢方面	$<1\times10^{-100}$	精氨琥珀酸盐合成酶
Wernicke-korsakoff 综合征；*TKT*	新陈代谢方面	$<1\times10^{-40}$	转羟乙醛酶
Variegate pophyria；*PPOX*	新陈代谢方面	$<1\times10^{-40}$	原卟啉氧化酶
青年发病的成年型糖尿病(MODY2)；*GCK*	新陈代谢方面	$<1\times10^{-40}$	Hxk1，己糖激酶
吉特曼综合征；*SLC12A3*	新陈代谢方面	$<1\times10^{-40}$	CCC Na-K-Cl 转运子
胱氨酸尿 1 型；*SLC3A1*	新陈代谢方面	$<1\times10^{-40}$	α-葡(萄)糖苷酶
胆囊纤维化；*ABCC7*	新陈代谢方面	$<1\times10^{-40}$	ABC 转运子
Bartter 综合征；*SLC12A1*	新陈代谢方面	$<1\times10^{-40}$	CCC Na-K-Cl 转运子
Menkes 综合征；*ATP7A*	神经方面	$<1\times10^{-100}$	P 型铜 ATP 酶
遗传性耳聋；*MYO15*	神经方面	$<1\times10^{-100}$	Myo51 Ⅴ类肌浆球蛋白
Zellweger 综合征；*PEX1*	神经方面	$<1\times10^{-40}$	AAA-家族 ATP 酶
Thomsen 疾病；*CLCN1*	神经方面	$<1\times10^{-40}$	ClC 氯化物通道蛋白
脊髓小脑性共济失调 6 型(SCA6)；*CACNA1A*	神经方面	$<1\times10^{-40}$	VIC 钠通道蛋白
肌强直性肌萎症；*DM1*	神经方面	$<1\times10^{-40}$	Orb6 丝氨酸/苏氨酸蛋白激酶
马科恩-亚白特症候群；*GNAA1*	神经方面	$<1\times10^{-40}$	Gpa1 GNP
lowe 眼脑肾综合征；*OCRL*	神经方面	$<1\times10^{-40}$	PIP 磷酸化酶
dents；*CLCN5*	神经方面	$<1\times10^{-40}$	ClC 氯化物通道蛋白
Coffin-lowry 综合征；*RPS6KA3*	神经方面	$<1\times10^{-40}$	丝氨酸/苏氨酸蛋白激酶
Agelman 综合征；*UBE3A*	神经方面	$<1\times10^{-40}$	泛素蛋白 lgase
肌萎缩性脊髓侧索硬化症；*SOD1*	神经方面	$<1\times10^{-40}$	Sod1，氧化物歧化酶
Oguschi2 型；*RHKIN*	神经方面	$<1\times10^{-40}$	丝氨酸/苏氨酸蛋白激酶
家族心脏疾病；*MYH7*	心脏方面	$<1\times10^{-100}$	Myo2，Ⅱ型肌浆球蛋白
肾小管酸中毒；*ATP6B1*	肾脏方面	$<1\times10^{-100}$	Ⅴ型 ATP 酶

注：1. 图中的得分是用 BLAST 工具搜索得到的 *E* 值，表中 $<1\times10^{-40}$ 的指分数在 1×10^{-40} 和 1×10^{-100} 之间。
2. 资料来自 Wood 等（2002），授权使用。

正如我们在第 3 章中所讨论的那样，PAM 和 BLOSUM 矩阵（图 3-10～图 3-17）显示了不同氨基酸的进化替代相对速率是不同的。谷氨酸通常会改变为天门冬氨酸（在 PAM250 矩阵中，分数为+2）；这两个残基都是酸性的因此分享共同的生物化学特征。但是，谷氨酸很少会改变为赖氨酸（在 PAM250 矩阵中，分数为 0），并且从谷氨酸到赖氨酸的突变通常会导致人类疾病。Miller 和 Kumar（2001）把从真核生物中统计的氨基酸改变频率（图 18-18 中用圆圈表示）和从病人中统计的氨基酸的改变频率（图 18-18 中用方块表示）用表的形式表示出来。

> Jackson 实验室网页"人类疾病的老鼠模型：老鼠和人基因的同源性"的网址为 http://jaxmice.jax.org/jaxmicedb/html/model_975.shtml/。

这些分析表明和疾病相关的氨基酸改变倾向于在保守的残基中发生，并且，在人类疾病中发现的氨基酸改变并不经常发生在对照的物种中。

对老鼠的研究。老鼠基因组由老鼠基因组测序联盟发布（Waterston 等，2002）。该基因组发布以后，我们拥有了最重要的关于人类疾病的动物模型，并且可以得到许多重要的资源。

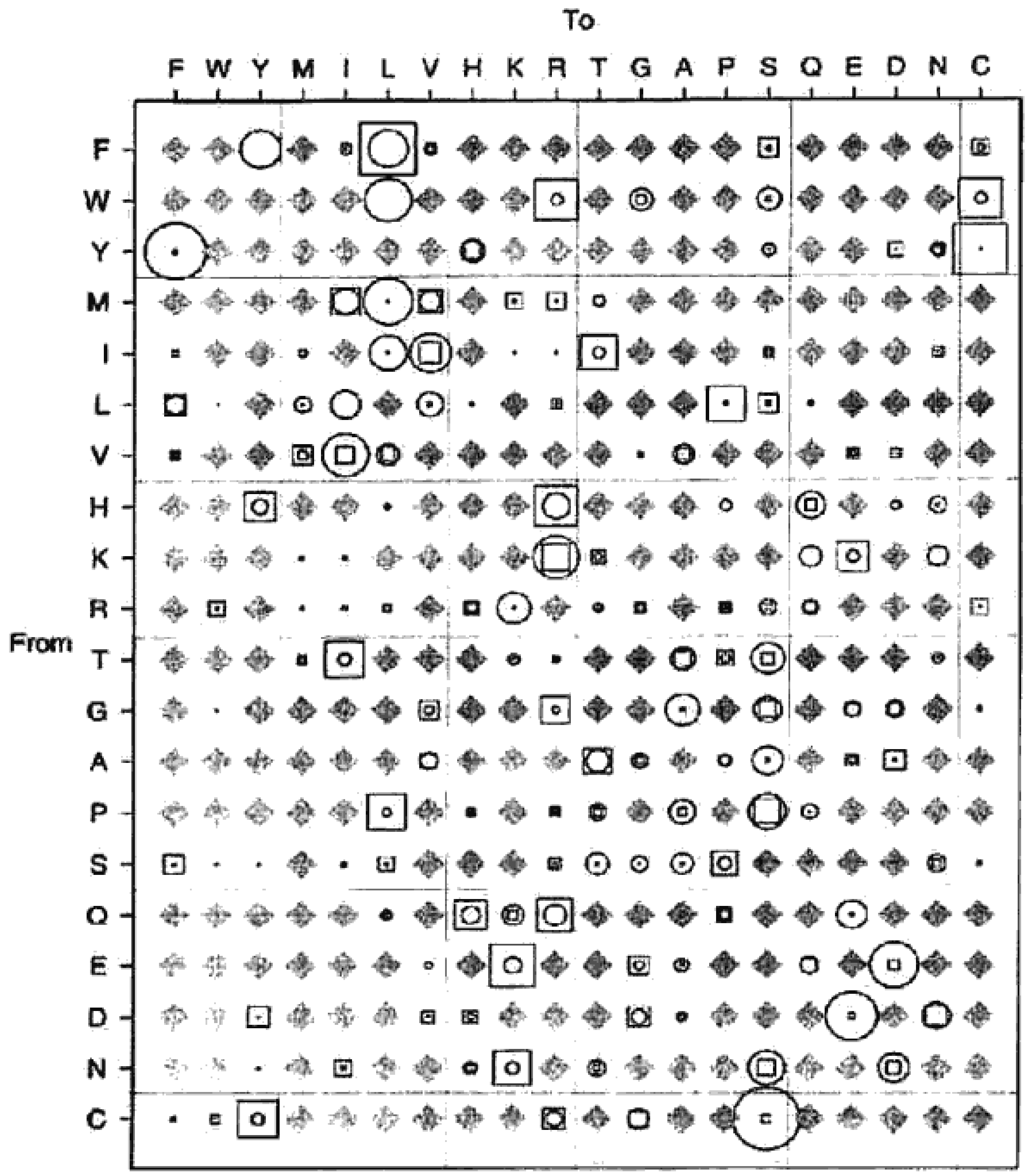


图 18-18 发生在人类疾病中的氨基酸替换通常不根据自然选择原则。该图是关于氨基酸之间转换概率的图，显示了从各种真核生物的比较中得到的氨基酸相互之间转换的概率（圆圈），以及病人中的氨基酸相互之间转换的概率（方框）。符号的大小表示了给定氨基酸转换概率的大小。钻石状的符号表示现在还不能得到单个碱基突变的转换概率。图来自Miller和Kumar（2001），授权使用

① FANTOM数据库，该数据库包含RIKEN老鼠基因组计划的部分数据，其中有老鼠cDNA克隆全长的信息（Bono等，2002），以及相关的描述老鼠全长克隆和人类疾病基因相关信息的数据库（表18-17）。

表 18-17 老鼠 cDNA 全长克隆和相应人类疾病基因

项目		FANTOM2 克隆	
		是	否
老鼠同源基因	是	740	173
	否	67	13

注：来自 FANTOM2 数据库，见 http://fantom2.gsc.riken.go.jp/supplement/disease-genes/。

② Jackson实验室的网站列出了老鼠和人的同源基因，并提供了研究人类疾病的老鼠模型。

③ 高效的诱导有机体突变的物质［如氮-乙烷基-氮-亚硝基脲（ENU）］或放射物已经用到老鼠身上从而得到人类疾病的模型（Hrabe de Angelis和Strivens，2001）。Nolan等（2002）对这种方法进行了详细的讨论，并制作了网站，介绍了如何进行高通量突变实验以及如何找到新表型的基因突变。

④ 人和老鼠同源基因数据库Dysmorphic包含了对基因异常表达的描述以及人和老鼠的联结图谱。

⑤ 整个老鼠的目录描述了人类疾病的老鼠模型。

人和老鼠同源基因数据库 Dysmorphic 由伦敦大学开发，网址为 http://www.hgmp.mrc.ac.uk/DHMHD/dysmorph.html。
你可以通过以下网站获得关于这些老鼠的信息：http://www.rodentia.com/wmc/domain_genome.html#transgenics 和 http://www.rodentia.com/wmc/domain_mouse.html。

对老鼠基因组的测序主要是由 Celera 公司和一个公共组织（第 16 章）完成的。Celera 公司测得了一些纯老鼠株系的基因组 DNA，并且标明了老鼠对传染性疾病（表 18-18）和复杂遗传疾病（表 18-19）感染的难易程度。比较基因组的数据将有可能解释为什么纯老鼠株系在不同疾病的感染上会不同。Adrian Hill（2001）对人类疾病被感染的难易程度的遗传学和基因组学机制做了综述。

表 18-18 传染疾病对受损老鼠的感染性

传染疾病	纯老鼠株系	
	A/J	C57BL/6J
军团肺炎	易感染	不易感染
疟疾	易感染	不易感染
病毒性肝炎	不易感染	易感染
鼠科艾滋病	不易感染	易感染

注：来自 http://www.celera.com/genomics/other/mousec16pres/home.cfm? ppage=5&returnLoc=index.jsp。

表 18-19 普通的复合物病对老鼠株的易感性

复合物疾病	纯老鼠株系	
	A/J	C57BL/6J
关节炎	易感染	不易感染
Colon 癌	易感染	不易感染
肺癌	易感染	不易感染
哮喘	易感染	不易感染
动脉硬化症	不易感染	易感染
高血压	不易感染	易感染
Ⅱ型糖尿病	不易感染	易感染
骨质疏松症	不易感染	不易感染
肥胖症	不易感染	易感染

注：资源来自 http://www.celera.com/genomics/other/mousec16pres/home.cfm? ppage = 5&returnLoc = index.jsp。

有关研究机构测得了含有与人类 687 种疾病基因同源的老鼠的基因组序列（Waterston 等，2002）。令人惊奇的是，在与这 687 种人类疾病基因同源的基因中，有几十个野生型老鼠的基因和人类疾病基因完全相同。这些基因在表 18-20 中列出。这表明，老鼠自身可能并不感染人类与这些基因有关的疾病，因此我们对是否能够得到人类这些疾病的老鼠模型要慎重考虑。由上我们也不难想像到，老鼠有些旁系同源基因在人类中没有出现，因此它们才不会感染人类的某些疾病。同样，纯老鼠株系放置在不同的环境中比野生型的老鼠更容易紧张，并且它们对疾病的易感性也会不同。

表 18-20 与人类疾病相关的基因序列变量完全匹配的老鼠序列

疾 病	OMIM	突 变
赫氏病	142623	E251K
消除白质的白血病	603896	R113H
IVA 型黏多糖(储积)病	253000	R376Q
乳腺癌	113705	L829S
	600185	V211A, Q2421H
帕金森病	601508	A53T
结节性脑硬化	605284	Q654E
6 型巴比二氏综合征	209900	T57A
间皮瘤	156240	N93S
长 QT 综合征 5	176261	V109I
胆囊纤维化	602421	F87L, V754M
卟啉多元化	176200	Q127H
非霍奇金淋巴瘤	605027	A25T, P183L
免疫缺失疾病	102700	R142Q
2D 型肌萎缩性疾病	254110	P30L
长链酰基辅酶 A 脱氢酶缺失	201460	Q333K
1B 型尤塞综合征	276902	G955S
慢性非球形红细胞溶血性贫血	206400	A295V
Mantle 细胞淋巴瘤	208900	N750K
轻型肌营养不良症	300377	H2921R
激素完全不敏感综合征	300068	G491S
前列腺癌	176807	P269S, S647N
节段性回肠炎	266600	W157R

注：改编自 Waterston 等（2002），授权使用。

18.10 人类疾病研究组织

我们已经给生物信息学下了定义，即利用计算算法和数据库来研究基因、蛋白和基因组的科学。对于人类疾病来说，许多数据库可以从万维网上得到（Guttmacher，2001）。表 18-21 是这些资源的列表，其中包括提供了几百种不同疾病家族信息的组织。

表 18-21 研究人类疾病的基本网络资源

研究定位	描 述	网 址
疾病、基因异常和相关主题	卡罗琳斯卡医学院（斯德哥尔摩）	http://www.mic.ki.se/Diseases/
遗传异常概率数据库(FIDD)	威尔士大学医学院遗传医学研究中心	http://archive.uwcm.ac.uk/uwcm/mg/fidd/
基因信息描述	人类基因数据库，包括基因产物和它们涉及的疾病	http://bioinfo.weizmann.ac.il/cards/
基因和疾病(NCBI)	Chromosome（一个组织）提供了对 60 多个对疾病的描述	http://www.ncbi.nlm.nih.gov/disease/
基因治疗	西雅图华盛顿大学基因治疗信息资源	http://www.geneclinics.org/
遗传联盟	国际上个人、专家和基因支持组织的联合	http://www.geneticalliance.org/; http://www.geneticalliance.org/diseaseinfo/search.html
通过定位克隆鉴别遗传疾病基因	在 NIH 的国际人类基因组研究中心(NHGRI)	http://genome.nhgri.nih.gov/clone

续表

研究定位	描述	网址
对残疾儿童和青年研究的国际信息中心	美国信息和推介中心	http://www.nichcy.org/
研究罕见基因异常信息的国际组织(NORD)	致力于帮助患有罕见疾病病人和为这些病人服务的组织的私人健康组织联合会	http://www.rarediseasc.org/
男性孟德尔遗传在线(OMIM)	超过12000个条目	http://www.ncbi.nlm.nih.gov/centrez/query.fcgi? db=OMIM

18.11 疾病基因的功能分类

我们对人类疾病的研究包括研究人类疾病的原理。人类疾病的多样性是很广泛的，因此生物信息学领域可以帮助我们了解疾病的逻辑关系。Jimenez-Sanchez 等（2001）分析了和人类疾病有关的923个人类基因。这些基因主要导致单性生殖混乱，正如我们最近才知道的相对很少的几个基因突变就可以使这个复杂的系统混乱一样。他们根据这些疾病基因的蛋白产物的功能对它们进行聚类［图 18-19(a)］。其中酶是最大的功能类并且该类中有31%的基因产物都和人类疾病有关。与此相比，仅有15%的位点克隆疾病基因编码酶蛋白。因此，认为是酶缺失而导致疾病的基因突变的认识可能存在某些偏见。

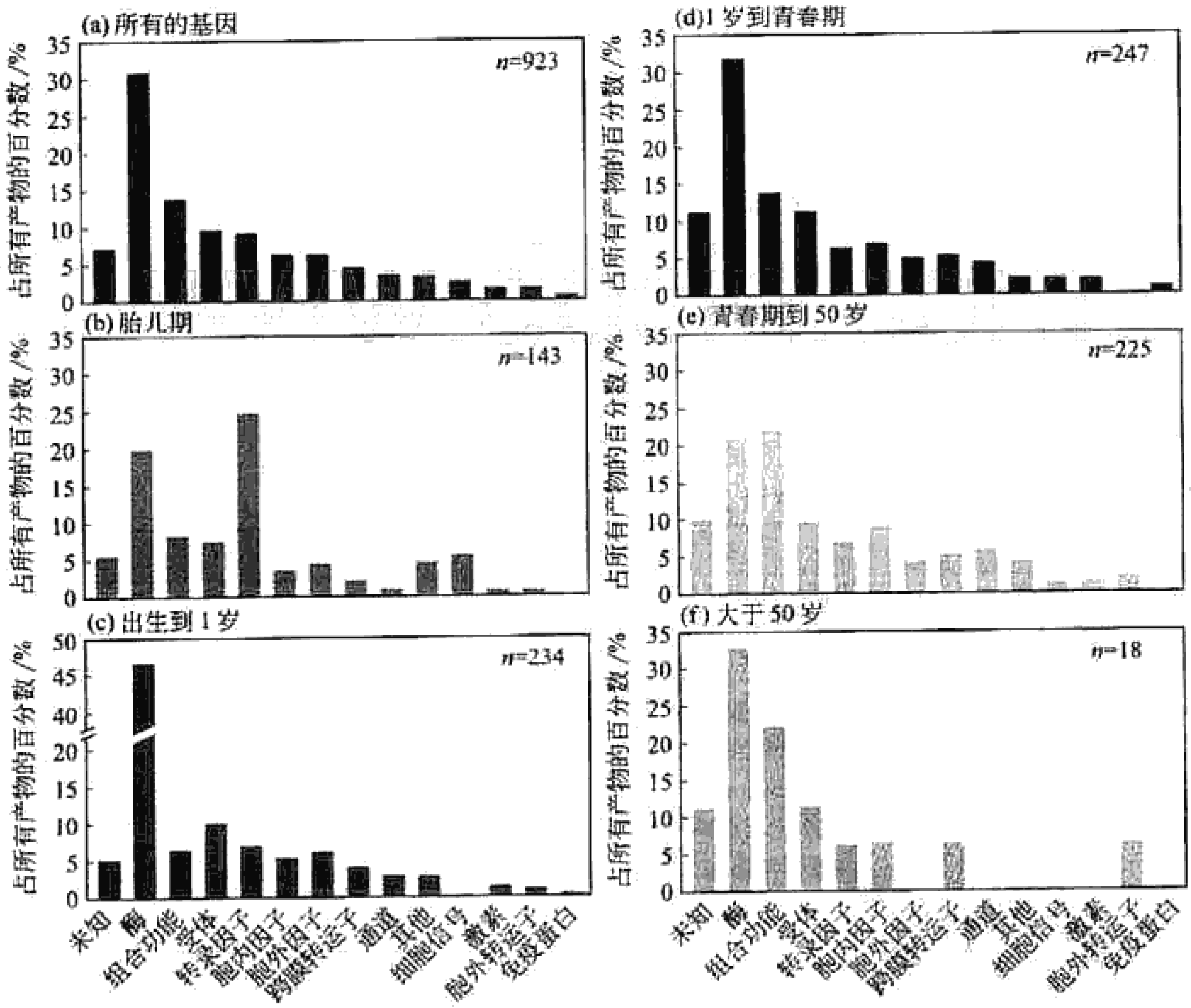


图 18-19 疾病相关基因蛋白产物的功能（选自 Jimenez-Sanchez 等，2001）：(a) 所有的基因；(b)～(f) 根据疾病发生的典型年龄分出的疾病基因。授权使用

Jimenez-Sanchez 等（2001）进一步分析了基因产物的功能和疾病发病时间的关系［图 18-19(b)～图 18-19(f)］。基因编码酶蛋白和转录因子与子宫内疾病的发生有很大关系，表明转录因子在早期发育中有很重要的作用。酶涉及的从婴儿到青春期人的部分疾病见图 18-19(b)～图 18-19(d)。胎儿和它母亲的代谢系统相连，因此即使它有一个基因发生缺失，也有存活的能力。当出生后，这种疾病就会显现出来。如果疾病基因编码的酶蛋白在疾病中作用不明显，那么这种疾病就会晚发作［图 18-19(e)］。

当分析疾病的四个功能类（发病概率、遗传模式、发病年限和减少生命时间）的任何一类时（图 18-20，最左边列）发现该样本所有疾病的发病概率都很低。发病概率很低这个事件表明疾病基因群体（包含单个基因异常现象）普遍能够被研究。遗传模式主要倾向于在隐性常染色体上发生，尤其是编码酶蛋白的基因。正如图 18-19 描述的一样，发病时间和胚胎内的转录因子有关，编码酶蛋白的基因涉及的疾病发病时间为从出生到 1 岁，对于受体基因涉及的疾病来说发病年龄为从 1 岁到青春期和成年，对于具有调控功能的蛋白（例如具有稳定、激活和折叠其他蛋白的蛋白）涉及的疾病的发病时间一般为青年期。而对于疾病减少人

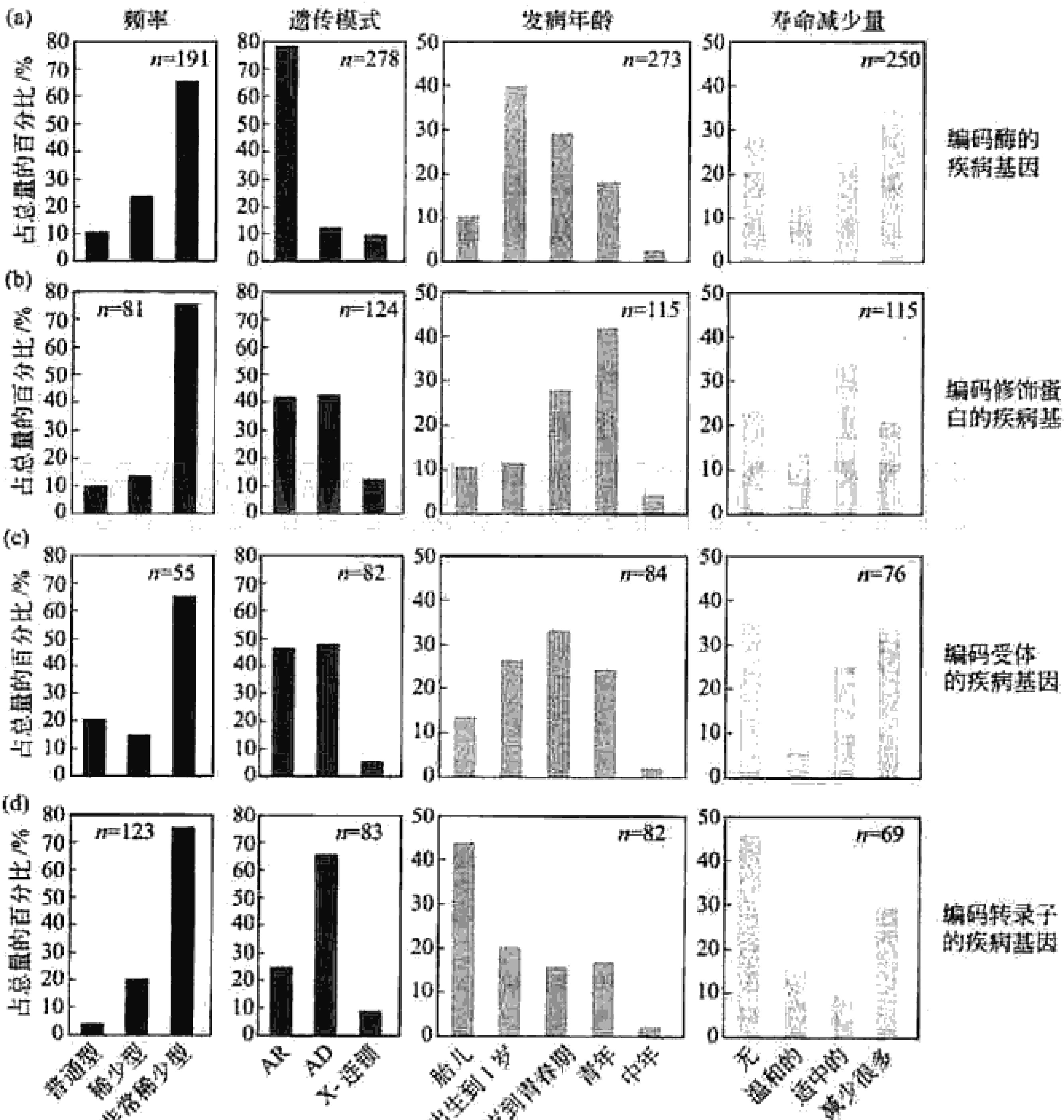


图 18-20 根据相关基因编码蛋白的功能得到的疾病的特征。AR 表示隐性常染色体；AD 表示显性常染色体；青年人指从青春期到小于 50 岁的人；老年人指大于 50 岁的人。图选自 Jimenez-Sanchez 等（2001），授权使用

的寿命方面，同一功能类的疾病基因并没有很统一的模式，它们减少人寿命的能力不同。

这些研究是对于研究人类疾病逻辑性关系的早期探索。这些基因组研究的成果会随着我们对基因组信息的逐渐了解而进一步充实。对于功能的分析需要依靠生物信息学和基因组学工具的帮助，从而阐明基因型和疾病表型之间的关系。

18.12 展望

有很多类型的生物信息学方法可以应用到人类疾病研究中，这些方法包括以下3个方面。

① 人类疾病是因为DNA序列改变的结果。这些改变可以在分子序列数据库（例如GenBank）中找到。

② 人类疾病数据库有一个主要的任务就是整理关于疾病相关基因的信息。这些数据库中，有一个比较重要的中心数据库OMIM以及特异位点突变数据库。

③ 功能基因组镜像为研究相关基因和疾病进程之间的机制提供了依据。

18.13 常见问题

我们研究功能图谱的目的就是让突变基因的基因型和疾病的表型相关联。我们可以用光谱方法处理疾病。从对疾病的表型研究开始，我们要了解有哪些疾病基因，这些基因在什么时候突变，这些突变会导致哪些疾病发生。从研究基因开始，我们关心的是在该基因突变时会导致何种疾病发生。现在，在技术手段上我们把从疾病研究基因和从基因研究疾病这两个问题结合起来几乎不可能。对于多数疾病来说，疾病相关基因的发现并不能导致新的处理或治理方式的发现，或者也不能使人们弄明白该疾病的病理生理学。例如对于肌肉营养不良症和蕾特症，我们希望通过生物信息学和功能基因组学的方法弄明白分子水平上的疾病发生的生化途径，但是现在却没有办法实现。对于上面的问题，我们也许可以通过利用模式生物研究致病基因的功能的方式来解决，或者也可以通过高通量技术来解决，例如那些能够检测敏感细胞对变异基因的转录响应的基因芯片技术。

18.14 网络资源

相关网络资源见表18-22。

表 18-22 研究癌症的网络资源

资源	描述	网址
ACOR	与癌症相关的网络资源	http://www.acor.org/
血液病学和肿瘤学方面的遗传学和细胞遗传学图集	关于在线杂志和癌症数据库的综述	http://www.infobiogen.fr/services/chromcancer/
癌症染色体异常工程	对癌症中染色体的选择和其特征的描述	http://cgap.nci.nih.gov/Chromosomes/CCAP
癌症基因注释工程(CGAP)	在NCBI的网站中	http://www.ncbi.nlm.nih.gov/ncicgap/
癌症基因组计划	Sanger研究中心	http://www.sanger.ac.uk/CCP/
关于癌症的网络	国际癌症研究中心	http://cancernet.nci.nih.gov/
关于癌症的网页	癌症资源站点	http://cancerweb.ncl.ac.uk/
关于儿童癌症的网页	儿童癌症资源列表	http://www.CancerIndex.org/

18.15 问题讨论

1. 许多神经性疾病（如蕾特症、亨廷顿疾病）等对脑功能有破坏性。对于这样的疾病中的一部分疾病来说，它们对应的疾病基因在单细胞生物（如真菌）中有同源基因，你对此问题是怎么想的？

2. 现在，芯片技术是怎样用来研究人类疾病的？何种癌症样本已经得到了？

18.16 问题

1. *WT1* 是 Wilms 癌症的癌抑制基因并且对肾的发育也很重要。在这个基因中是否有一些 SNPs？并且它们是否对蛋白序列有选择性？*WT1* 编码蛋白突变区的分布是怎样的？

(a) 进入 LocusLink 并且搜索 *WT1*。点击网页中的紫色栏，就可以进入含有 *WT1* 基因的 SNPs 网页。

(b) 检测 SNPs，看看是否有些 SNPs 能改变蛋白序列。

(c) 试着“比对 SNP”从而找到更多的 *WT1* 的 SNPs。进入网站 http://www.ncbi.nlm.nih.gov/SNP/并点击左面栏中的“Blast SNP”。你可以对搜索进行限制，可以仅搜索第 11 条染色体（*WT1* 所在的位置）。

(d) 在 OMIM 中探询 *WT1*。

—该基因位于哪个染色体上？

—大约多少等位点已经知道？

—如果想研究 Wilm 癌症，你是否能够得到这种病人基因异常的细胞系？（提示：点击“Coriell”）。

(e) 在 *WT1* 中寻找未知的突变。进入 OMIM 中关于 *WT1* 的网页并点击 HGMD（人类基因突变数据库）。在这个网站中有各种分类（错义、无义、缺失等）的所有突变。

—关于这个基因，它的突变导致了多少明显的表型？

—基于全局的突变图谱（见 HGMD 网页上面栏的链接），在该基因中的突变是否倾向于聚集或者均匀分布。

(f) 在第 11 号染色体上 *WT1* 上游的 3 个疾病基因是什么？而在它下游的 3 个疾病基因又是什么？*WT1* 在第 11 号染色体的短臂还是长臂上？利用 OMIM 来解决这些问题。

(g) 利用 NCBI 的人类染色体图谱数据库来寻找 *WT1* 上游和下游的 3 个基因。而该数据库显示了已知的所有基因（并不只是疾病相关基因）。你可以利用两种办法来浏览 NCBI 网站：

- 从 OMIM 网站中进入 *WT1*，然后点击“location”。
- 进入 NCBI 主页，然后从主页右侧页面选择“human map viewer”。

2. 有多少遗传疾病拥有已知和它们相关的序列？

(a) 在 LocuLink 网站上寻找“disease _ known AND has _ seq.”，并且限制搜索对象为人类（通过自动弹出的按钮选择）。这样就可以找到异常的基因并根据该基因的和序列数据发现疾病。

(b) 进入 OMIM 网站并且查看对这个问题的回答的统计结果（在网页左侧栏目中）。

18.17 自测题

1. 在人类中，遗传造成的异常基因符合孟德尔遗传定律的占人类总疾病基因的百分数为

(a) 1%；

(b) 10%；

(c) 50%；

(d) 不能准确确定所占的百分数

2. 根据现有知识，由于环境导致的基因异常（如中毒等造成的疾病）是由单个基因决定的

(a) 正确；

(b) 错误

3. 下面关于单个基因异常的描述哪一个是最好的？

(a) 单个基因异常是因为单个基因的突变造成的。它们与复合物异常相比是疾病中的一个基本类；

(b) 单个基因异常主要是由于在单个基因内部的突变造成的，不过疾病的发生总是由于多个基因的共同作用造成的，因此它们和复合物异常都是代表疾病的同一类；

(c) 单个基因异常主要是由于单个基因内部的突变造成的，并且在这个过程中，突变总会是同义置换；

(d) 单个基因异常主要是由于单个基因内部的突变造成的，并且在这个过程中，突变总会是反义置换

4. 总的来说，罕见疾病在美国影响的人数为

(a) 200000；

(b) 200 万；

(c) 2500 万；

(d) 1 亿

5. 单基因异常趋向于

(a) 在普通的人群中比较少见，并且发病时间比较早；

(b) 在普通的人群中比较常见，并且发病时间比较早；

(c) 在普通的人群中比较少见，并且发病时间比较晚；

(d) 在普通的人群中比较常见，并且发病时间比较晚

6. 位点特异性数据库有几百个，它们提供的信息在中心数据库（如 OMIM 和 GeneCards）中没有的是

(a) 综合描述基因在疾病中的作用；

(b) 全面列出和疾病相关的突变；

(c) 链接众多基金会和其他的一些组织的网站；

(d) 链接染色体图谱显示导致疾病的基因

7. 染色体异常包括单体型、三体型和基因重组。基因组芯片可以用来检测这些改变，那么这些芯片的组成为

(a) 定位在染色体上的寡聚核苷酸，它们之间的间隔 1～10Mb；

(b) cDNA 克隆，并且用来了解染色体位点；

(c) BAC 克隆，并且用来了解染色体位点；

(d) FISH 指针

8. 人类疾病基因和许多物种的基因都有同源性，这些物种包括蠕虫、昆虫和真菌。许多人类的蛋白都和疾病有关，现在已经可以对同源蛋白的多序列进行比对。这些都表明在人类蛋白质中氨基酸的位置和疾病发生相关，这表明这些和疾病有关的残基

(a) 在其他的物种中非常保守；

(b) 在其他的物种中有些保守；

(c) 在其他物种中有很弱的保守性；

(d) 仅仅有时候会发生在同源序列中

18.18 推荐读物

研究人类疾病的基本资料就是《基于遗传疾病的代谢和分子》(Scriver 等，2001)，该书有四卷（含几百章），并从各个方面（如孟德尔异常、复合物异常、疾病的逻辑问题、突变机制和动物模型）介绍人类疾病。另外推荐的解释疾病的是由 Barton Childs 和 David Valle (2000) 撰写的一个评论，它在《遗传学和人类遗传学年鉴》的第一卷。

2001 年 2 月出版的《科学》和《自然》杂志包含了由 Leena Peltonen、Victor Mckusick (2001) 及 Jimenez-Sanchez 等 (2001) 撰写的关于人类疾病的文章。

这些文章都很详细和客观地介绍了人类疾病方面的知识和最新的研究，读者如果感兴趣，不妨找来读读，必会大有裨益。

参考文献

Ahringer, J. Turn to the worm! *Curr. Opin. Genet. Dev.* **7**, 410-415 (1997).

Allaman-Pillet, N., Djemai, A., Bonny, C., and Schorderet, D. F. Methylation status of CpG sites and methyl-CpG binding proteins are involved in the promoter regulation of the mouse Xist gene. *Gene Expr.* **7**, 61-73 (1998).

Amir, R. E., et al. Rett syndrome is caused by mutations in X-linked *MECP2*, encoding methyl-CpG-binding protein 2. *Nat. Genet.* **23**, 185-188 (1999).

Amir, R. E., and Zoghbi, H. Y. Rett syndrome: Methyl-CpG-binding protein 2 mutations and phenotype-genotype correlations. *Am. J. Med. Genet.* **97**, 147-152 (2000).

Anderson, R. N. Deaths: Leading causes for 1999. *Natl. Vital. Stat. Rep.* **49**, 1-87 (2001).

Antonarakis, S. E. Recommendations for a nomenclature system for human gene mutations. Nomenclature Working Group. *Hum. Mutat.* **11**, 1-3 (1998).

The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815 (2000).

Armstrong, D. D. Review of Rett syndrome. *J. Neuropathol. Exp. Neurol.* **56**, 843-849 (1997).

Bailey, A., et al. Autism as a strongly genetic disorder: Evidence from a British twin study. *Psychol. Med.* **25**, 63-77 (1995).

Bailey, A., Phillips, W., and Rutter, M. Autism: Towards an integration of clinical, genetic, neuropsychological, and neurobiological perspectives. *J. Child Psychol. Psychiatry* **37**, 89-126 (1996).

Baker, P., Piven, J., Schwartz, S., and Patil, S. Brief report: Duplication of chromosome 15q11-13 in two individuals with autistic disorder. *J. Autism Dev. Disord.* **24**, 529-535 (1994).

Barton, M., Volkmar, F. How commonly are known medical conditions associated with autism? *J. Autism Dev. Disord.* **28**, 273-278, (1998).

Bauman, M. L., Kemper, T. L., and Arin, D. M. Microscopic observations of the brain in Rett syndrome. *Neuropediatrics* **26**, 105-108 (1995).

Beaudet, A. L., Scriver, C. R., Sly, W. S., and Valle, D. Genetics, biochemistry, and molecular bases of variant human phenotypes. In Scriver et al. (Eds.), *The Metabolic & Molecular Bases of Inherited Disease*, Vol. 1. Mc Graw-Hill, New York, 2001, pp. 3-45.

Belichenko, P. V., Hagberg, B., and Dahlstrom, A. Morphological study of neocortical areas in Rett syndrome. *Acta Neuropathol. (Berl.)* **93**, 50-61 (1997).

Beroud, C., Collod-Beroud, G., Boileau, C., Soussi, T., and Junien, C. UMD (Universal mutation database): a generic software to build and analyze locus-specific databases. *Hum. Mutat.* **15**, 86-94, 2000.

Bienvenu, T., et al. *MECP2* mutations account for most cases of typical forms of Rett syndrome. *Hum. Mol. Genet.* **9**, 1377-1384 (2000).

Bono, H., Kasukawa, T., Furuno, M., Hayashizaki, Y., and Okazaki, Y. FANTOM DB: Database of Functional Annotation of RIKEN Mouse cDNA Clones. *Nucleic Acids Res.* **30**, 116-118 (2002).

Brown, A. F., and McKie, M. A. MuStaR and other software for locus-specific mutation databases. *Hum. Mutat.* **15**, 76-85 (2000).

Buetow, K. H., Edmonson, M. N., and Cassidy, A. B. Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet.* **21**, 323-325 (1999).

Bundey, S., Hardy, C., Vickers, S., Kilpatrick, M. W., and Corbett, J. A. Duplication of the 15q11-13 region in a patient with autism, epilepsy and ataxia. *Dev. Med. Child. Neurol.* **36**, 736-742 (1994).

Cedar, H. DNA methylation and gene activity. *Cell* **53**, 3-4 (1988).

Cheadle, J. P., et al. Long-read sequence analysis of the *MECP2* gene in Rett syndrome patients: Correlation of disease severity with mutation type and location. *Hum. Mol. Genet.* **9**, 1119-1129 (2000).

Childs, B., and Valle, D. *Genetics, Biology and Disease*. Annual Reviews, Palo Alto, CA, 2000, pp. 1-19.

Ciaranello, A. L., and Ciaranello, R. D. The neurobiology of infantile autism. *Annu. Rev. Neurosci.* **18**, 101-128 (1995).

Claustres, M., Horaitis, O., Vanevski, M., and Cotton, R. G. Time for a unified system of mutation description and reporting: A review of locus-specific mutation databases. *Genome Res.* **12**, 680-688 (2002).

Colantuoni, C., Purcell, A. E., Bouton, C. M., and Pevsner, J. High throughput analysis of gene expression in the human brain. *J. Neurosci. Res.* **59**, 1-10 (2000).

Collins, F. S., Brooks, L. D., and Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8**, 1229–1231, (1998).

Cook, E. H., Jr. et al. Autism or atypical autism in maternally but not paternally derived proximal 15q duplication. *Am. J. Hum. Genet.* **60**, 928–934 (1997).

Cook, E. H., Jr. et al. Linkage-disequilibrium mapping of autistic disorder, with 15q11-13 markers. *Am. J. Hum. Genet.* **62**, 1077–1083 (1998).

Cowles, C. R., Joel, N. H., Altshuler, D., and Lander, E. S. Detection of regulatory variation in mouse genes. *Nat. Genet.* **32**, 432–437 (2002).

den Dunnen, J. T., and Antonarakis, S. E. Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. *Hum. Mutat.* **15**, 7–12 (2000).

DiMauro, S., and Schon, E. A. Mitochondrial DNA mutations in human disease. *Am. J. Med. Genet.* **106**, 18–26 (2001).

Dipple, K. M., and McCabe, E. R. Modifier genes convert "simple" Mendelian disorders to complex traits. *Mol. Genet. Metab.* **71**, 43–50, (2000).

Dipple, K. M., Phelan, J. K., and McCabe, E. R. Consequences of complexity within biological networks: Robustness and health, or vulnerability and disease. *Mol. Genet. Metab.* **74**, 45–50 (2001).

Flejter, W. L., et al. Cytogenetic and molecular analysis of inv dup(15) chromosomes observed in two patients with autistic disorder and mental retardation. *Am. J. Med. Genet.* **61**, 182–187 (1996).

Fombonne, E. The epidemiology of autism: A review. *Psychol. Med.* **29**, 769–786 (1999).

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).

Garrod, A. E. *Inborn errors of metabolism: The Croonian Lectures delivered before the Royal College of Physicians of London, in June, 1908.* Frowde, Hodder and Stoughton, London (1909).

Garrod, A. E. *Inborn factors in disease: An essay.* Clarendon Press, Oxford (1931).

Garrod, A. E. The incidence of alkaptonuria: A study in chemical individuality. *Lancet* **ii**, 1616–1620 (1902).

Geerdink, N., et al. *MECP2* mutation in a boy with severe neonatal encephalopathy: Clinical, neuropathological and molecular findings. *Neuropediatrics* **33**, 33–36 (2002).

Gillberg, C., and Wing, L. Autism: Not an extremely rare disorder. *Acta Psychiatr. Scand.* **99**, 399–406 (1999).

Goodstadt, L., and Ponting, C. P. Sequence variation and disease in the wake of the draft human genome. *Hum. Mol. Genet.* **10**, 2209–2214 (2001).

Guttmacher, A. E. Human genetics on the web. *Annu. Rev. Genomics Hum. Genet.* **2**, 213–233 (2001).

Hagberg, B., Aicardi, J., Dias, K., and Ramos, O. A progressive syndrome of autism, dementia, ataxia, and loss of purposeful hand use in girls: Rett's syndrome: report of 35 cases. *Ann. Neurol.* **14**, 471–479 (1983).

Hammer, S., Dorrani, N., Dragich, J., Kudo, S., and Schanen, C. The phenotypic consequences of *MECP2* mutations extend beyond Rett syndrome. *Ment. Retard. Dev. Disabil. Res. Rev.* **8**, 94–98 (2002).

Hamosh, A., Scott, A. F., Amberger, J., Valle, D., and McKusick, V. A. Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.* **15**, 57–61 (2000).

Hamosh, A., et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **30**, 52–55 (2002).

Hermsen, M. A., Weiss, M. M., Meijer, G. A., and Baak, J. P. Detection of chromosomal abnormalities by comparative genomic hybridization. *Methods Mol. Biol.* **175**, 47–55 (2001).

Hill, A. V., The genomics and genetics of human infectious disease susceptibility. *Annu. Rev. Genomics Hum. Genet.* **2**, 373–400 (2001).

Holt, I. J., Harding, A. E., and Morgan-Hughes, J. A. Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies. *Nature* **331**, 717–719, 1988.

Hrabe de Angelis, M., and Strivens, M. Large-scale production of mouse phenotypes: The search for animal models for inherited diseases in humans. *Brief. Bioinform.* **2**, 170–180 (2001).

Huppke, P., Laccone, F., Kramer, N., Engel, W., and Hanefeld, F. Rett syndrome: Analysis of *MECP2* and clinical characterization of 31 patients. *Hum. Mol. Genet.* **9**, 1369–1375 (2000).

International Molecular Genetic Study of Autism Consortium. A full genome screen for autism with evidence for linkage to a region on chromosome 7q. *Hum. Mol. Genet.* **7**, 571–578 (1998).

Jellinger, K., Armstrong, D., Zoghbi, H. Y., and Percy, A. K. Neuropathology of Rett syndrome. *Acta Neuropathol.* **76**, 142–158 (1988).

Jimenez-Sanchez, G., Childs, B., and Valle, D. Human disease genes. *Nature* **409**, 853–855 (2001).

Kiberstis, P., and Roberts, L. It's not just the genes. *Science* **296**, 685 (2002).

Kim, S. J., and Cook, E. H., Jr. Novel de novo nonsense mutation of *MECP2* in a patient with Rett syndrome. *Hum. Mutat.* **15**, 382–383 (2000).

Knoppers, B. M., and Laberge, C. M. Ethical guideposts for allelic variation databases. *Hum. Mutat.* **15**, 30–35 (2000).

Krawczak, M., and Cooper, D. N. The human gene mutation database. *Trends Genet.* **13**, 121–122 (1997).

Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22**, 139–144 (1999).

Kruglyak, L., and Nickerson, D. A. Variation is the spice of life. *Nat Genet.* **27**, 234–236 (2001).

Lamb, J. A., Moore, J., Bailey, A., and Monaco, A. P. Autism: Recent molecular genetic advances. *Hum. Mol. Genet.* **9**, 861–868 (2000).

Lichter, P., Joos, S., Bentz, M., and Lampel, S. Comparative genomic hybridization: Uses and limitations. *Semin. Hematol.* **37**, 348–357 (2000).

Lupski, J. R. Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**, 417–422 (1998).

Miller, M. P., and Kumar, S. Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet.* **10**, 2319–2328 (2001).

Murray, C. J. L., and Lopez, A. D. (eds.) *The Global Burden of Disease.* Harvard University Press, Cambridge (1996).

Naidu, S. Rett syndrome: A disorder affecting early brain growth. *Ann. Neurol.* **42,** 3–10 (1997).

Nan, X., Campoy, F. J., and Bird, A. MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. *Cell* **88,** 471–481 (1997).

Nass, S., and Nass, M. M. K. Intramitochondrial fibers with DNA characteristics. *J. Cell Biol.* **19,** 613–629 (1963).

Ng, H. H., and Bird, A. DNA methylation and chromatin modification. *Curr. Opin. Genet. Dev.* **9,** 158–163 (1999).

Nolan, P. M., Hugill, A., and Cox, R. D. ENU mutagenesis in the mouse: Application to human genetic disease. *Briefings Functional Genomics Proteomics* **1,** 278–289 (2002).

Olsson, I., Steffenburg, S., and Gillberg, C. Epilepsy in autism and autisticlike conditions. A population-based study. *Arch. Neurol.* **45,** 666–668 (1988).

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stem, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P., and Cox, D. R. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294,** 1719–1723 (2001).

Pauling, L., Itano, H. A., Singer, S. J., and Wells, I. C. Sickle cell anemia, a molecular disease. *Science* **110,** 543–548 (1949).

Peltonen, L., and McKusick, V. A. Genomics and medicine. Dissecting human disease in the postgenomic era. *Science* **291,** 1224–1229 (2001).

Philippe, A., et al. Genome-wide scan for autism susceptibility genes. Paris Autism Research International Sibpair Study. *Hum. Mol. Genet.* **8,** 805–812 (1999).

Phillips, M. S., Lawrence, R., Sachidanandam, R., Morris, A. P., Balding, D. J., Donaldson, M. A., Studebaker, J. F., Ankener, W. M., Alfisi, S. V., Kuo, F. S., Camisa, A. L., Pazorov, V., Scott, K. E., Carey, B. J., Faith, J., Katari, G., Bhatti, H. A., Cyr, J. M., Derohannessian, V., Elosua, C., Forman, A. M., Grecco, N. M., Hock, C. R., Kuebler, J. M., Lathrop, J. A., Mockler, M. A., Nachtman, E. P., Restine, S. L., Varde, S. A., Hozza, M. J., Gelfand, C. A., Broxholme, J., Abecasis, G. R., Boyce-Jacino, M. T., and Cardon, L. R. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat Genet.* **33,** 382–387 (2003).

Pinkel, D., et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20,** 207–211 (1998).

Piven, J. The biological basis of autism. *Curr. Opin. Neurobiol.* **7,** 708–712 (1997).

Pollack, J. R., et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23,** 41–46 (1999).

Purcell, A. E., Jeon, O. H., Zimmerman, A. W., Blue, M. E., and Pevsner, J. Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. *Neurology* **57,** 1618–1628 (2001).

Rapin, I. Autism. *N. Engl. J. Med.* **337,** 97–104 (1997).

Rapin, I., and Katzman, R. Neurobiology of autism. *Ann. Neurol.* **43,** 7–14 (1998).

Razin, A. CpG methylation, chromatin structure and gene silencing–a three-way connection. *EMBO J* **17,** 4905–4908 (1998).

Reiter, L. T., Potocki, L., Chien, S., Gribskov, M., and Bier, E. A systematic analysis of human disease-associated gene sequences in *Drosophila melanogaster*. *Genome Res.* **11,** 1114–1125 (2001).

Rossi, P. G., Parmeggiani, A., Bach, V., Santucci, M., and Visconti, P. EEG features and epilepsy in patients with autism. *Brain Dev.* **17,** 169–174 (1995).

Rubin, G. M., et al. Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).

Safran, M., et al. GeneCards(TM) 2002: Towards a complete, object-oriented, human gene compendium. *Bioinformatics* **18,** 1542–1543 (2002).

Schon, E. A. Mitochondrial genetics and disease. *Trends Biochem. Sci.* **25,** 555–560 (2000).

Scriver, C., Beaudet, A., Sly, W., and Valle, D. (Eds.). *The Metabolic and Molecular Basis of Inherited Disease.* McGraw-Hill, New York (2001).

Scriver, C. R., and Childs, B. *Garrod's Inborn Factors in Disease.* New York, Oxford University Press (1989).

Scriver, C. R., Nowacki, P. M., and Lehvaslaiho, H. Guidelines and recommendations for content, structure, and deployment of mutation databases. *Hum. Mutat.* **13,** 344–350 (1999).

Scriver, C. R., Nowacki, P. M., and Lehvaslaiho, H. Guidelines and recommendations for content, structure, and deployment of mutation databases: II. Journey in progress. *Hum. Mutat.* **15,** 13–15 (2000).

Sherry, S. T., Ward, M., and Sirotkin, K. Use of molecular variation in the NCBI dbSNP database. *Hum. Mutat.* **15,** 68–75 (2000).

Singer, C. *The Fasciculo di Medicina Venice 1493.* R. Lier and Co., Florence, 1925.

Smalley, S. L., Asarnow, R. F., and Spence, M. A. Autism and genetics. A decade of research. *Arch. Gen. Psychiatry* **45,** 953–961 (1988).

Szatmari, P., Jones, M. B., Zwaigenbaum, L., and MacLean, J. E. Genetics of autism: Overview and new directions. *J. Autism Dev. Disord.* **28,** 351–368 (1998).

Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L., and Kwok, P. Y. Overlapping genomic sequences: A treasure trove of single-nucleotide polymorphisms. *Genome Res.* **8,** 748–754 (1998).

Thomas, C. L. (ed.) *Taber's Cyclopedic Medical Dictionary.* F. A. Davis Company, Philadelphia, 1997.

Todd, J. A. Multifactorial diseases: Ancient gene polymorphism at quantitative trait loci and a legacy of survival during our evolution. In C. Scrivner et al. (Eds.), *The Metabolic & Molecular Bases of Inherited Disease,* Vol. 1. Mc Graw-Hill, New York, 2001, pp. 193–201.

Turner, M., Barnby, G., and Bailey, A. Genetic clues to the biological basis of autism. *Mol. Med. Today* **6,** 238–244 (2000).

Van den Veyver, I. B., and Zoghbi, H. Y. Methyl-CpG-binding protein 2 mutations in Rett syndrome. *Curr. Opin. Genet. Dev.* **10,** 275–279 (2000).

Volkmar, F. Recently diagnosed with autism, autism or not. *J. Autism Dev. Disord.* **28,** 269–270 (1998).

Volkmar, F. R., and Nelson, D. S. Seizure disorders in autism. *J. Am. Acad. Child. Adolesc. Psychiatry* **29,** 127–129 (1990).

Wallace, D. C., Singh, G., Lott, M. T., Hodge, J. A., Schurr, T. G., Lezza, A. M., Elsas, L. J. 2nd, Nikoskelainen, E. K. Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. *Science* **242,** 1427–1430, 1988a.

Wallace, D. C., Zheng, X. X., Lott, M. T., Shoffner, J. M., Hodge, J. A., Kelley, R. I., Epstein, C. M., Hopkins, L. C. Familial mitochondrial encephalomyopathy (MERRF): genetic, pathophysiological, and biochemical characterization of a mitochondrial DNA disease: *Cell* **55**, 601–610, 1988b.

Wan, M., et al. Rett syndrome and beyond: Recurrent spontaneous and familial *MECP2* mutations at CpG hotspots. *Am. J. Hum. Genet.* **65**, 1520–1529 (1999).

Wang, Z., and Moult, J. SNPs, protein structure, and disease. *Hum. Mutat.* **17**, 263–270 (2001).

Waterhouse, L. Genes tPA, Fyn, and FAK in autism? *J. Autism Dev. Disord.* **27**, 220–223 (1997).

Waterston, R. H., et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

Wood, V., et al. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871–880 (2002).

Xiang, F., et al. Mutation screening in Rett syndrome patients. *J. Med. Genet.* **37**, 250–255 (2000).

Zeev, B. B., et al. Rett syndrome: Clinical manifestations in males with *MECP2* mutations. *J. Child Neurol.* **17**, 20–24 (2002).

Zuckerlandl, E., and Pauling, L. Molecular disease, evolution, and genic heterogeneity. In M. Kasha, and B. Pullman, (Eds.), *Horizons in Biochemistry*, Albert Szent-Gyorgyi Dedicatory Volume. Academic Press, New York (1962).

后 记

生物学已经进入了基因组时代，测序和对基因组的分析为生物学引入了新的方法。本书的前 3 章着重讲述如何获取信息，尤其是 DNA 和蛋白质的序列数据。然后，我们讲述了序列的两两比对和多序列比对。在本书的中间 3 章，我们讨论了基因表达谱和蛋白质以及蛋白质组学的问题，其中还解释了分子进化。最后，在本书的最后 3 章中，我们通过生命树探讨了基因组问题，并且对关于人类疾病的研究做了阐述。

我们正处于人类基因组计划已经完成，并且其他很多物种的序列正在测序的生物科学时代。许多在国际人类基因组研究中心的科学家（包括 Francis Collins、Eric Green、Alan Guttmacher 和 Mark Guyer）最近表达了他们对基因组学下一个时代的看法（Collins 等，2003）。经过几百个科学家两年来的讨论，他们列出了未来基因组学的 15 个挑战性问题。这些挑战性问题主要分为 3 个方面：基因组学和生物学、基因组学和人类健康，以及基因组学和社会学。

你可以更详细地阅读他们提出的关于基因组学的 15 个挑战性的问题，Collins 等（2003）的文章可以从网站 http://www.genome.gov/About/Planning/ 中找到。

① 基因组学和生物学。

a. 全面确定人类基因组中关于结构和功能的编码元件。

b. 阐明遗传网络和蛋白通路以及确定它们如何控制细胞和组织的表型问题。

c. 更进一步了解人类基因组中可遗传的性状特征问题。

d. 弄清楚物种的进化多样性以及内在的机制问题。

e. 开发新的技术使在科学研究和临床应用方面获取基因组信息更便捷。

② 基因组学和人类健康。

a. 发展用于检测遗传因素对疾病和药物响应的作用的鲁棒策略。

b. 发展新的技术应用于鉴别对健康有利且抵抗疾病的基因。

c. 发展新的基于基因组的方法来预测人类对疾病的易感性和药物反应、对疾病的早期诊断，以及对疾病不同阶段的分子分型的确定。

d. 利用对基因和通路的新的理解来开发新的对疾病治疗的有效方法。

e. 分析遗传危害信息是怎样干扰临床治疗的，它们又对人类健康和活动有怎样的影响，以及它们怎样导致人类疾病。

f. 开发基于基因组的工具来提高对疾病的治疗效率，改善人类健康。

③ 基因组学和社会学

a. 开发新的技术使基因组信息应用到医疗和非医疗方面。

b. 弄清楚基因组学与种族和种族划分之间有无关系，以及这些关系之间的因果联系。

c. 弄清楚基因组与人类特征及其活动能力的关系。

d. 如何利用基因组的信息来对种族边缘进行确定。

所有的这些目标都很大，完成它们中的任何一个都需要真正弄清楚生物信息学和基因组学的工具。我们希望本书中对分子序列、数据库、算法、基因组、生物学原理等的探讨能够对致力于研究生物学的学生和学者有所帮助，并且帮助他们朝上面提到的那些宏伟的目标前进。

参考文献

Collins, F. S., Green, E. D., Guttmacher, A. E., and Guyer, M. S. A vision for the future of genomics research. *Nature* **422**, 835-847, 2003.

附　录

GCG 软件包：用于蛋白和 DNA 分析

第一部分：GCG 简介

GCG（Genetics Computer Group）软件包集合了 130 多个相关的软件程序，用于蛋白、RNA、DNA 序列的分析。GCG 是商业产品，其应用的范围同 Vector NTI（Informax 公司产品）以及基于 UNIX 系统的 Emboss 免费工具相似（http://www.hgmp.mrc.ac.uk/Software/Emboss/）。

有关 GCG 的详细资料可以访问网站 http://www.gcg.com（或者 http://www.accelrys.com/about/gcg.html）。GCG 于 1982 年在美国的威斯康星-麦迪逊大学遗传系建立。目前得到全世界 650 多个研究机构 40000 多位科学家的广泛使用。

GCG 涵盖的内容如下：

比较分析	导入导出
数据库搜索	作图
DNA/RNA 二级结构	引物挑选
编辑发布	蛋白分析
进化	SeqLab
片段装配	翻译
基因查找和模式识别	Utilities

第二部分：一般命令

UNIX 一般命令

当使用 GCG 程序包时是在 UNIX 环境下运行，所以下面介绍一些有用的命令：

＞cat x	显示文档 x
＞more x	显示文档 x，以一页一页的方式（按空格键前进到下一页）
＞ls	列出目录中的文件
＞ls *.pep	列出目录中所有以 .pep 后缀结尾的文件
＞∧C	中止目前运行的内容，回到命令行
＞∧S	暂停屏幕输出

>∧Q　　继续输出

>man　　在线帮助手册

GCG 一般命令

在 GCG 中，有一些很重要的命令需要掌握。

启动帮助

对于所使用的任何一个程序，查看帮助菜单非常有用。一般而言，帮助菜单提供以下一些内容：

① 纵览该程序功能；

② 举例说明如何使用；

③ 描述使用的算法以及相关参考文献；

④ 比较此程序和其他相关的 GCG 程序的功能。

启动帮助的命令：

>genhelp　　启动 GCG 内所有 130 个程序的帮助内容

>genhelp gap　　启动名为 GAP 程序的帮助内容

组织目录文件

当使用 GCG 时，会很快累积几十个或上百个 GCG 文件，一个很有效的组织这些文件的方法是，为每个你正在工作的项目建立一个单独的目录（见图 1）。建目录时，只要键入命令“mkdir”再在后面加新目录名即可。进入要工作的目录，键入“down a. dir”命令。可以通过敲击命令“over any. dir”在目录之间横向切换。要回到主起始目录，可以敲击“home”。无论何时如你想确认自己所处的位置，只要敲击“ls”就可列出所拥有的文件以及当前目录或子目录的名称。对组织文件的命令总结如下：

>mkdir　　建立新目录

>mkdir a. dir　　建立名为 a. dir 的新目录

>down a. dir　　从 home 目录向下切换到 a. dir

>over b. dir　　切换至名为 b. dir 的目录

>up　　向上切换至更高层的目录

>home　　回到主 home 目录

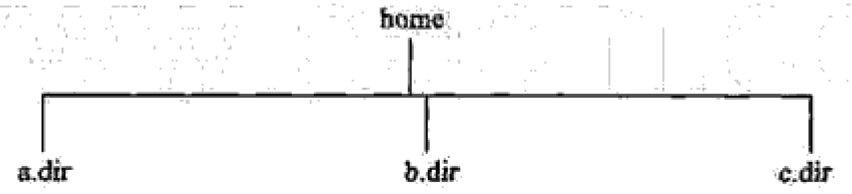


图 1

命名文件

一个很方便的命名蛋白文件的方法是在后面添加“. pep”后缀（如 actin. pep），命名 DNA 文件则添加“. seq”后缀（如 actin. seq）。当在主序列编辑器 seqed 中首次创建一个文件时，必须命名该文件。在 GCG 程序中，形如 a. pep 或 a. seq 的名字可用于确认输入程序的数据适当（例如，. seq 文件用于限制酶分析，. pep 文件用于疏水性分析）。然而还是有些例外，例如，翻转 DNA 序列显示底链序列时，应该添加后缀 . rev。

删除、重命名、复制文件

当删除文件时，千万要小心操作。因为删除容易，而恢复却不容易。键入“>rm x”删除名为 x 的文件（rm 代表 remove）：

>rm x　　删除名为 x 的文件

>rm *z*　　删除所有名字中有 z 的文件

>rm *. pep　　删除所有含有后缀 . pep 的肽或蛋白文件

重命名文件，使用命令 mv（move）

>mv somename. pep bettername. pep　　重命名文件 somename. pep 为 bettername. pep

复制文件，使用命令 cp（copy）

>cp *. pep~/somedirectory/　复制所有含 . pep 后缀的文件到 somedirectory 目录下

注意符号～代表 home 目录。

第三部分：输入和编辑序列

GCG 的主要特征是其操作序列的能力。大多数情况下，需要一开始导入序列至 GCG 中，有关导入内容的典型例子包括以下内容：

① 从 Core Facility 得到某 DNA 序列并输入至电脑中，想知道这条序列和已知 DNA 序列或蛋白序列之间的匹配程度如何；

② 直接测得一条蛋白的序列，想要分析它的性质；

③ 某人给你发送一条重组融合蛋白，你想分析它的性质；

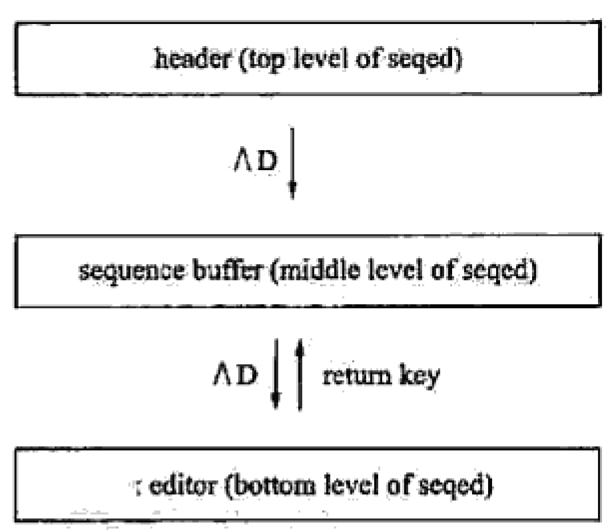
④ 想分析两条蛋白是否很同源。

许多时候，你会在 NCBI（National Center for Biotechnology Information）或 SwissProt（见第 2 章）中找到开始要用的蛋白或 DNA 序列。

复制 DNA 或蛋白序列，然后用序列编辑器 seqed 粘贴到 GCG 中：

>seqed x. pep　　建立（或编辑）名为 x. pep 的蛋白文件

>seqed y. seq　　建立（或编辑）名为 y. seq 的 DNA 文件



打开 seqed 程序，在页面上有三层：顶部的标题，中间的序列缓冲器，底部的命令行编辑器。敲击“∧D”（control D）向下移动至序列缓冲器，然后粘贴你的序列。

常用的编辑命令有：

：exit　　结束 seqed 进程，保存改动

：quit　　结束 seqed 进程，不保存改动

：include　粘贴另一条 DNA 序列

：help　　启动 seqed 的帮助

图 2 范例显示了在 seqed 程序中三层的情况。

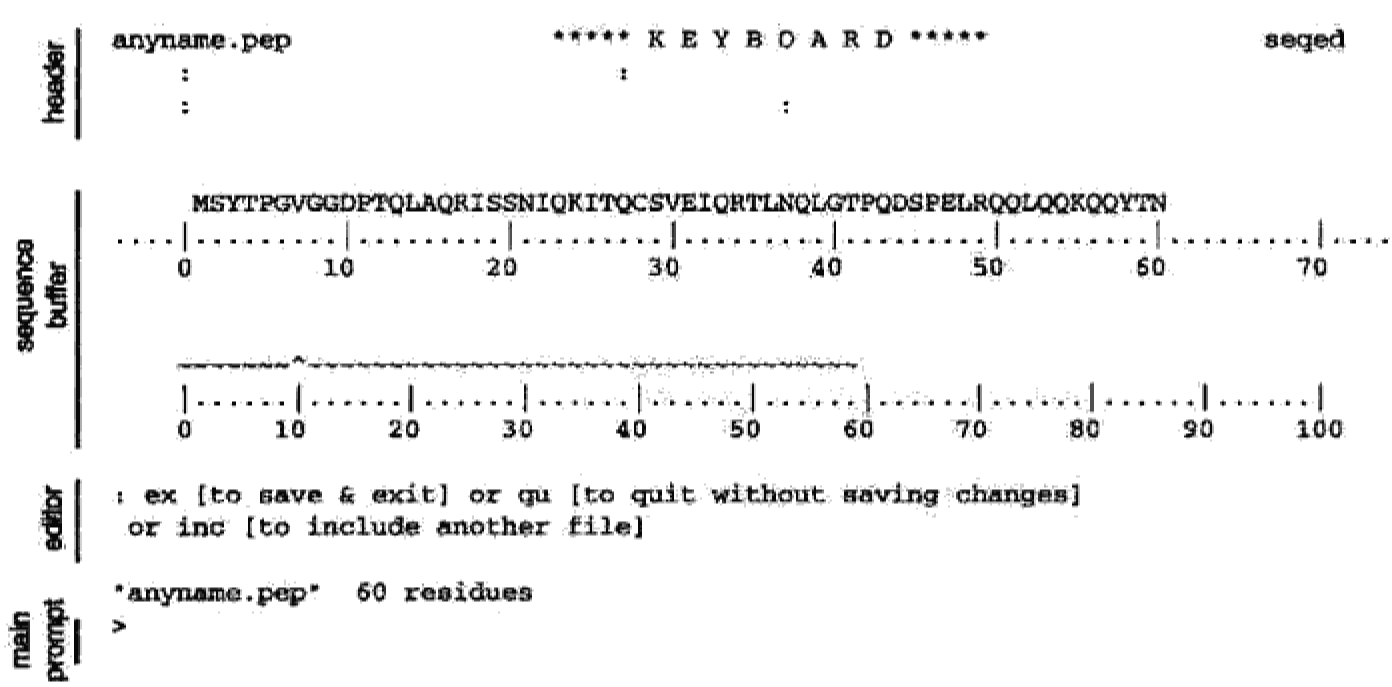


图 2

输入人视黄醇结合蛋白（RBP4，retinal-binding protein）至 GCG，敲入>seqed hsrbp. pep<return>（hsrbp 代表 homo sapiens retinol-binding protein）。注意 GCG 指针现在顶部标题区。

```
hsrbp.pep               ***** K E Y B O A R D *****            seqed
     :                                                              :
     :                                                              :
     :                                                              :
     :                                                              :
```

在浏览器（如 Netscape）中，打开 GenBank 里的 RBP 蛋白记录。复制 AC 号和其他相关的信息至标题区。完成后，按∧D进入主序列缓冲器。复制粘贴 RBP 的氨基酸序列至 GCG（注意 GCG 会忽略蛋白或核酸序列字母下的数字）：

```
test.pep                 ***** K E Y B O A R D *****                   seqed
    : LOCUS      NP_006735     199 aa                      PRI       08-JUN0 :
    : DEFINITION retinol-binding protein 4, interstitial precursor [Homo . :
    : ACCESSION  NP_006735
    : PID        g5803139

crllnldgtcadsysfvfsrdpnglppeaqkivrqrqeelclarqyrlivhngycdgrsernll
....|.........|.........|.........|.........|.........|.........|.........|....
  140       150       160       170       180       190       200       210
    ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~^
    |......|......|......|......|......|......|......|......|......|......|
    0     20     40     60     80    100    120    140    160    180    200
```

下一步，在 GCG 中，按∧D从序列缓冲器切换至命令行。命令行有一冒号（“:”）。敲“ex”（即“exit”）。

```
: ex

You have successfully created the file hsrbp.pep.
Type > ls to confirm that this file has appeared in your directory.
Type > cat hsrbp.pep to see this file displayed:

> cat hsrbp.pep
!!AA_SEQUENCE 1.0
LOCUS        NP_006735 199 AA PRI 08-JUN-2000
DEFINITION   RETINOL-BINDING PROTEIN 4, INTERSTITIAL PRECURSOR
             [HOMO SAPIENS].
ACCESSION    NP_006735
PID          G5803139
VERSION      NP_006735.1 GI:5803139
DBSOURCE     REFSEQ: ACCESSION NM_006744.2
KEYWORDS     .
SOURCE       HUMAN.
ORGANISM     HOMO SAPIENS

HSRBP.PEP LENGTH: 199  AUGUST 30, 2000 19:19  TYPE: P  CHECK: 8315

       1 MKWVWALLLL AAWAAAERDC RVSSFRVKEN FDKARFSGTW YAMAKKDPEG

      51 LFLQDNIVAE FSVDETGQMS ATAKGRVRLL NNWDVCADMV GTFTDTEDPA

     101 KFKMKYWGVA SFLQKGNDDH WIVDTDYDTY AVQYSCRLLN LDGTCADSYS

     151 FVFSRDPNGL PPEAQKIVRQ RQEELCLARQ YRLIVHNGYC DGRSERNLL
>
```

Optionally, you can use the following commands:

> onecase	make sequences all upper or lowercase
> onecase *.pep	make all protein sequences one case

可使用下面任一个命令：

>onecase　　　　序列全部大写或小写

>onecase *.pep　　所有蛋白序列大写或小写

在附录中，我们也使用 syntaxins 的例子。此种蛋白结合囊泡蛋白，协调囊泡对接到靶膜上。

第四部分：两序列比对

BestFit　　比较两个蛋白或 DNA 序列，局部比对

GAP　　比较两个蛋白或 DNA 序列，全局比对

TFastA　　比较一个蛋白同 DNA 编码的 6 种可能的蛋白

BestFit 使用举例

在本例中，比较两个蛋白（人 syntaxin 1a 和人 syntaxin 7），方法是敲入“bestfit”和待比较的蛋白名称（之间留一空格）。

接着，使用 cat 命令查看比对结果。

```
> bestfit humsyn1a.pep humsyn7.pep

BestFit makes an optimal alignment of the best segment of similarity
between two sequences. Optimal alignments are found by inserting gaps
to maximize the number of matches using the local homology algorithm
of Smith and Waterman.

                  Begin (* 1 *) ?
                End (*   289 *) ?

                  Begin (* 1 *) ?
                End (*   261 *) ?

What is the gap creation penalty (* 12 *) ?

What is the gap extension penalty (* 4 *) ?

What should I call the paired output display file (* humsyn1a.pair *) ?

Aligning .............-.
Aligning .............-..

         Gaps:       2
      Quality:     194
Quality Ratio:   0.911
 % Similarity:  39.151
       Length:     216


 > cat humsyn1a.pair
  BESTFIT of:  humsyn1a.pep  check:  5551  from:  1  to:  289

 LOCUS        HUMSYN1A      2088 bp    mRNA          PRI 23-NOV-1994
 DEFINITION   Human syntaxin 1A mRNA, complete cds.
 ACCESSION    L37792
 NID          g577487
 KEYWORDS     syntaxin 1A.
 SOURCE       Homo sapiens male adult brain cDNA to mRNA . . .

  to: humsyn7.pep  check: 338  from: 1  to: 261
```

```
Symbol comparison table: /prog/GCG/gcgcore/data/rundata/blosum62.cmp
CompCheck: 6430

         Gap Weight:     12          Average Match:  2.912
      Length Weight:      4       Average Mismatch: -2.003
            Quality:    194                 Length:    216
              Ratio:  0.911                   Gaps:      2
 Percent Similarity: 39.151       Percent Identity: 26.887

      Match display thresholds for the alignment(s):
                  | = IDENTITY
                  : =    2
                  . =    1

humsyn1a.pep x humsyn7.pep October 28, 1998 19:27  ..

           .         .         .         .         .
  71 TKEELEELMSDIKKTANKVRSKLKSIEQSIEQEEGLNRSSADLRIRKTQH 120
     | ::  ||   :..         |  :. |.:   |  . .: | || |
  41 TPQDSPELRQQLQQKQQYTNQLAKETDKYIKEFGSLPTTPSEQRQRKIQK 90

              .         .         .         .         .
 121 STLSRKFVEVMSEYNATQSDYRERCK...GRIQRQLEITGRTTTSEELED 167
       |  .    :. :   |    || |    |:.    :.|        |
  91 DRLVAECTTSLTNFQKVQRQAAEREKEFVARVRASSRVSGSFPEDSSKER 140

              .         .         .         .         .
 168 MLESGNPAIFASGIIMDSSISKQALSEIETRHSEIIKLENSIRELHDMFM 217
      | |          : |  |..  |  |  | | | .||  | ::.:.|
 141 NLVSWESQTQPQVQVQDEEITEDDLRLIHERESSIRQLEADIMDINEIFK 190

              .         .         .         .         .
 218 DMAMLVESQGEMIDRIEYNVEHAVDYVERAVSDTKKAVKYQSKARRKKIM 267
     |: |::  ||:.|| || |||.|  :|:.|     :|  || |. || :
 191 DLGMMIHEQGDVIDSIEANVENAEVHVQQANQQLSRAADYQRKS.RKTLC 239

              .
 268 IIICCVILGIVIASTV 283
     |||  .::|: | | :
 240 IIILNLVIGVAIISLI 255
```

输出结果注意事项：

① 比对采用单字母氨基酸编码，记住！

② 线表示氨基酸相同

③ 点表示氨基酸相似

④ gap 权重和长度罚分是可调节的参数，以此来优化比对结果。例如，如果你认为两个蛋白比较远源，比对结果不好，可以试着降低 gap 罚分，也许比对标准降低会产生比较好的比对结果（代价是加入更多的 gap）。

GCG 关于 BestFit 程序的文档（>genhelp bestfit）

Bestfit 对两序列相似度最好的部分做最优比对。最优比对采用 Smith-Waterman 局部同源算法，通过插入 gap 来最大化匹配的数目（见第 3 章）。

描述　BestFit 通过插入 gap 在两个序列相似度最好的部分做最优比对，然后显示比对结果，格式同 GAP 的输出结果相似。序列的长度可以相差很大，只要有一小部分片段相似即可。例如，你可以找一条短 RNA 序列，然后将其和整个线粒体基因组比对。

寻找相似部分　在 wisconsin 软件包中，BestFit 是识别两条未知关联序列之间最高相似度区域的最强的工具。

GAP 程序使用举例

```
> gap humsyn1a.pep humsyn7.pep

Gap uses the algorithm of Needleman and Wunsch to find the alignment
of two complete sequences that maximizes the number of matches and
minimizes the number of gaps.

           Begin (* 1 *) ?
         End (*   289 *) ?

           Begin (* 1 *) ?
         End (*   261 *) ?

What is the gap creation penalty (* 12 *) ?

What is the gap extension penalty (* 4 *) ?

What should I call the paired output display file (* humsyn1a.pair *) ?

Aligning .............-.
Aligning .............-..

        Gaps:      2
     Quality:    169
Quality Ratio:  0.648
 % Similarity: 35.659
      Length:    292
```

接着，查看比对结果。

```
> cat humsyn1a.pair
 GAP of: humsyn1a.pep  check: 5551  from: 1  to: 289

LOCUS        HUMSYN1A     2088 bp    mRNA        PRI 23-NOV-1994
DEFINITION   Human syntaxin 1A mRNA, complete cds.
ACCESSION    L37792
NID          g577487
KEYWORDS     syntaxin 1A.
SOURCE       Homo sapiens male adult brain cDNA to mRNA . . .

 to: humsyn7.pep  check: 338  from: 1  to: 261

 Symbol comparison table: /prog/GCG/gcgcore/data/rundata/blosum62.cmp
 CompCheck: 6430

        Gap Weight:    12     Average Match:  2.912
     Length Weight:     4   Average Mismatch: -2.003

           Quality:   169            Length:    292
             Ratio: 0.648              Gaps:      2

 Percent Similarity: 35.659 Percent Identity: 23.643

        Match display thresholds for the alignment(s):
            | = IDENTITY
            : =   2
            . =   1

humsyn1a.pep x humsyn7.pep October 28, 1998 20:08  ..
```

```
          .         .         .         .         .
  1 MKDRTQELRTAKDSDDDDDVAVTVDRDRFMDEFFEQVEEIRGFIDKIAEN 50
                                 |        :      :|. |
  1 .............................MSYTPGVGGDPTQLAQRISSN 21
             .         .         .         .         .
 51 VEEVKRKHSAILASPNPDEKTKEELEELMSDIKKTANKVRSKLKSIEQSI 100
    ::.: .    |  . |    | ::  ||   :..         |  :. |
 22 IQKITQCSVEIQRTLN.QLGTPQDSPELRQQLQQKQQYTNQLAKETDKYI 70
             .         .         .         .         .
101 EQEEGLNRSSADLRIRKTQHSTLSRKFVEVMSEYNATQSDYRERCK...G 147
    .:   |  . .: | || |   |  .    :. :   |   ||  |
 71 KEFGSLPTTPSEQRQRKIQKDRLVAECTTSLTNFQKVQRQAAEREKEFVA 120
             .         .         .         .         .
148 RIQRQLEITGRTTTSEELEDMLESGNPAIFASGIIMDSSISKQALSEIET 197
    |:.    :.|        |  | |          : |  |..  |  |
121 RVRASSRVSGSFPEDSSKERNLVSWESQTQPQVQVQDEEITEDDLRLIHE 170
             .         .         .         .         .
198 RHSEIIKLENSIRELHDMFMDMAMLVESQGEMIDRIEYNVEHAVDYVERA 247
    | | | .||  | ::.:.| |: |::  ||:.|| || |||.|  :|:.|
171 RESSIRQLEADIMDINEIFKDLGMMIHEQGDVIDSIEANVENAEVHVQQA 220
             .         .         .         .         .
248 VSDTKKAVKYQSKARRKKIMIIICCVILGIVIASTVGGIFA* 289
         :|  || |.|:   .||:  ||   :|.  : |:
221 NQQLSRAADYQRKSRKTLCIIIILNLVIGVAIISLIIWGLNH. 261
```

两蛋白同源性统计显著吗?

对于局部比对，决定两条核酸或蛋白序列是否同源最严格的方法是衡量 *E* 值（详见第 4 章）。对于全局比对，统计显著性比较难以估计，一种方法是做随机化测试（见图 3-30）。GCG 中操作如下：

敲入 “>gap humsyn1a. pep humsyn7. pep－ran＝50.”

“－ran＝50” 命令让程序比较 humsyn1a. pep 蛋白和 humsyn7. pep 蛋白，得到一个打分，然后程序将 humsyn7. pep 氨基酸序列随机打乱，再和 humsyn1a. pep 比较，再打分。50 次这样的随机打乱 humsyn7. pep 序列并和 humsyn1a. pep 比较。程序输出除了 GAP 一般结果外，还有 50 次随机化打分的均值和标准差。输出结果举例如下：

```
         Gap Weight:      12          Average Match:   2.912
      Length Weight:       4       Average Mismatch:  -2.003

            Quality:     194                 Length:     216
              Ratio:   0.911                   Gaps:       2
 Percent Similarity:  39.151     Percent Identity:   26.887

 Average quality based on 50 randomizations: 35.2 +/- 4.0
```

50 次随机化平均打分的这些数据结果可以解答你的一个很重要的问题：我比较两条蛋白得到的打分值（在本例中是 $Q=194$），会比我随机化其中一条蛋白 50 次得到的打分值好得多吗（$Q=35.2+/-4.0$）？答案是 yes，因为检验比较中的 *Z* 分值比随机化分值多出不止 3 倍的标准差，即 $Z=(194-35.2)/4.0=39.7$。如果 *Z* 比 3 小，那么你的两条蛋白可能并不显著相关。

GAP（全局比对）和 **BestFit**（局部比对，类似于 BLAST）

注意，虽然 GAP 和 BestFit 相似，每个都可比对两条序列，但它们有明显的差别。BestFit 是做局部比对优化，在两条蛋白或核酸序列之间寻找最高同源区。GAP 是做全局比对优化，扫描整条序列。在上面

syntaxin 的例子里，BestFit 显示了更好的氨基酸 identity 的比值（27%，长度为 216 个氨基酸残基），而 GAP 只显示了 24%的比值，但长达 292 个氨基酸残基。

GCG 有关 GAP 的文档（>genhelp gap）

GAP 使用 Needleman-Wunsch 算法比对两条完整序列，使得匹配数最大，gap 数最少。

描述

GAP 考虑所有可能的比对方式和 gap 位置，产生拥有最多匹配碱基数和最少 gap 的比对方式。你可以在匹配碱基单元设置 gap 创建和延伸的罚分，也就是说，GAP 要在每个 gap 插入的位置扣除一定的分值（gap 创建罚分）。如果你的 gap 延伸罚分大于零，GAP 还必须扣除 gap 插入长度乘以 gap 延伸罚分的分值，GAP 使用的是 Needleman-Wunsch 算法（J Mol Biol，1970，48，443～453）。

TFastA 使用举例

```
> tfasta humsyn1a.pep humsyn7.seq

TFastA does a Pearson and Lipman search for similarity between a query
peptide sequence and any group of nucleotide sequences. TFastA translates
the nucleotide sequences in all six reading frames before performing the
comparison. It is designed to answer the question, "What implied peptide
sequences in a nucleotide sequence database are similar to my peptide
sequence?"

                Begin (* 1 *) ?
              End (*   289 *) ?

 What word size (* 2 *) ?

 Don't show scores whose E() value exceeds: (* 10.0 *):

 What should I call the output file (* humsyn1a.tfasta *) ?

          1 Sequences        1,673 aa searched
/user/pevsner/pevsner/syntaxin.
dir/humsyn7.seq

(Peptide) TFASTA of: humsyn1a.pep from: 1 to: 289 October 28, 1998 22:02

TO: humsyn7.seq Sequences:           1 Symbols:    1,673  Word Size: 2

Searching all six frames.
Scoring matrix: GenRunData:blosum50.cmp
Variable pamfactor used
Gap creation penalty: 16       Gap extension penalty: 4

The best scores are:                       frame init1  initn  opt
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq
!                                           (3)    171    189   268
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq
!                                           (2)     44     44    49
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq
!                                           (4)     38     38    41
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq
!                                           (1)     29     29    35
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq
!                                           (5)     29     29    29
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq
!                                           (6)     26     26    28
The list contains 6 entries.
How many alignments would you like to see (* 6 *) ?
```

[redacted]

```
Aligning...
 CPU time used:
       Database scan: 0:00: 0.1
Post-scan processing: 0:00: 0.1
      Total CPU time: 0:00: 0.3
 Output File: humsyn1a.tfasta

> cat humsyn1a.tfasta
!!SEQUENCE_LIST 1.0

(Peptide) TFASTA of: humsyn1a.pep from: 1 to: 289 October 28, 1998 22:02

 TO: humsyn7.seq Sequences:          1 Symbols:     1,673    Word Size: 2

 Searching all six frames.
 Scoring matrix: GenRunData:blosum50.cmp
 Variable pamfactor used
 Gap creation penalty: 16      Gap extension penalty: 4

The best scores are:                    frame init1  initn  opt..

/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq     Begin: 201  End: 917
!                                        (3)    171    189    268
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq     Begin: 557  End: 598
!                                        (2)     44     44     49
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq     Begin: 831  End: 920 Strand: -
!                                        (4)     38     38     41
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq     Begin: 310  End: 420
!                                        (1)     29     29     35
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq     Begin: 1619  End: 1645 Strand: -
!                                        (5)     29     29     29
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq     Begin: 1060  End: 1119 Strand: -
!                                        (6)     26     26     28
\\End of List
humsyn1a.pep
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq

SCORES  Frame:  (3)  Init1:   171 Initn:   189 Opt:   268
  25.3% identity  in  241 aa overlap

                   20        30        40        50        60         70
humsyn1a.pep DDDDVAVTVDRDRFMDEFFEQVEEIRGFIDKIAENVEEVKRKHSAILASPNPDEKTKEEL
                                           :|: |:::: :    |  : | :  | ::
humsyn7      LHSIXXNXRIRTSTMSYTPGVGGDPTQLAQRISSNIQKITQCSVEIQRTLN-QLGTPQDS
                   40        50        60        70        80         90

                   80        90       100       110       120        130
humsyn1a.pep EELMSDIKKTANKVRSKLKSIEQSIEQEEGLNRSSADLRIRKTQHSTLSRKFVEVMSEYN
              ||:::::   : : :  |  :: |::  :|  : :: | || |:: |  : :  :::::
humsyn7      PELRQQLQQKQQYTNQLAKETDKYIKEFGSLPTTPSEQRQRKIQKDRLVAECTTSLTNFQ
                  100       110       120       130       140        150

                  140          150       160       170       180        190
humsyn1a.pep ATQSDYRERCK---GRIQRQLEITGRTTTSEELEDMLESGNPAIFASGIIMDSSISKQAL
              :| :  || |   :|:: : :::|    :   |  | | :     :  ::|  |::: |
humsyn7      KVQRQAAEREKEFVARVRASSRVSGSFPEDSSKERNLVSWESQTQPQVQVQDEEITEDDL
                  160       170       180       190       200        210
```

[redacted]

```
                  200       210       220       230       240       250
humsyn1a.pep SEIETRHSEIIKLENSIRELHDMFMDMAMLVESQGEMIDRIEYNVEHAVDYVERAVSDTK
               |:  |:| |  :|| :| :::::| |::|::: ||::|| || |||:|  :|::| :: :
humsyn7      RLIHERESSIRQLEADIMDINEIFKDLGMMIHEQGDVIDSIEANVENAEVHVQQANQQLS
               220       230       240       250       260       270

                  260       270       280       289
humsyn1a.pep KAVKYQSKARRKKIMIIICCVILGIVIASTVGGIFAX
             :|: || |: || : |||  :::|::| | :
humsyn7      RAADYQRKS-RKTLCIIIILNLVIGVAIISLIIWGLNHXSYKGAHCRTTLSKLCRKIPVIM
               280       290       300          310    320       330

humsyn7      FFXLLFXSYCIKDGSHTLLFLLGGLGGFLWIKSDIFXYXKIFXMSLLTXLPWSSIXXLLS
               340       350       360          370    380       390

humsyn1a.pep
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq

SCORES   Frame: (2) Init1:    44 Initn:    44 Opt:    49
  50.0% identity in 14 aa overlap

                  30        40        50        60        70        80
humsyn1a.pep TVDRDRFMDEFFEQVEEIRGFIDKIAENVEEVKRKHSAILASPNPDEKTKEELEELMSDI
                                                   :::||:| ||::||
humsyn7      EGPEAGCXARERVCCSSKSQFQSVWQFSXGQLQRKESCILGKPNSTSSAGAGXRNYRGXP
               160       170       180       190       200       210

                  90       100       110       120       130       140
humsyn1a.pep KKTANKVRSKLKSIEQSIEQEEGLNRSSADLRIRKTQHSTLSRKFVEVMSEYNATQSDYR

humsyn7      PSYSXERIFYQATXSXYYGYXXNIXRFGNDDSXTRRCNRXHRSQCGKCRGARSASKSAAV
               220       230       240       250       260       270
humsyn1a.pep
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq

SCORES   Frame: (4) Init1:    38 Initn:    38 Opt:    41
  26.7% identity in 30 aa overlap

                  220       230       240       250       260       270
humsyn1a.pep HDMFMDMAMLVESQGEMIDRIEYNVEHAVDYVERAVSDTKKAVKYQSKARRKKIMIIICC
                                            | |    :::  || ::    : : :||||
humsyn7      KKHDYRNLPTXFRQCSATVCSFITSVVQSPYDETDNRNSNDKVKNDDAQGFSGFALIICC
                    230       240       250       260       270       280
                  280       289

humsyn1a.pep VILGIVIASTVGGIFAX

humsyn7      PXQLLICLLNVHLCIFHIGFYAIYYISLFMNHHSQIFKYFINIHNISFKLPDRRFSLMNK
                  290       300       310       320       330       340

humsyn1a.pep
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq

SCORES   Frame: (1) Init1:    29 Initn:    29 Opt:    35
  35.1% identity in 37 aa overlap

               80        90       100        110         120      130
humsyn1a.pep ELMSDIKKTANKVRSKLKSIEQSIEQEEGLNRSSADL-RIRKTQHSTLSRKFVEVMSEYN
                                             ||||: |  : | | |||:   :      |
humsyn7      RRSHSVLWKYKELXINLEHLKIHLNXGNSCNRSSSILTSLPKKQISTLKSLDLCPPPPVN
                    80        90       100        110         120      130
```

```
              140       150       160       170       180       190
humsyn1a.pep ATQSDYRERCKGRIQRQLEITGRTTTSEELEDMLESGNPAIFASGIIMDSSISKQALSEI
             :::: ||:
humsyn7      SVKGKYRRIAXWQSAQHHXQTSRRSRGRLLSERKSLLLEXEPVPECLAVFLRTAPKKGIL
                 140       150       160       170       180       190

humsyn1a.pep
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq

! CPU time used:
!         Database scan: 0:00:00.1
! Post-scan processing: 0:00:00.1
!        Total CPU time: 0:00:00.3
! Output File: humsyn1a.tfasta
```

何时使用 TFastA?

① 如果你有条 DNA 序列（例如，来自 BLAST 搜索或 core facility 酵母双杂交结果），TFastA 搜索会告诉你你的序列和任何已知序列的关系，如果因测序失误导致框移动，TFastA 搜索会显示出来，使用 GCG 的 map 程序可以使之得到修正。

② TFastA 是发现新基因的很有用的一个工具（此过程称为数据库挖掘，参见第 5 章）。若你在研究一个基因家族，想找到一个新成员，可到 GenBank 进行一系列 tblastn 的搜索，用你感兴趣的蛋白搜索表达序列标签数据库（EST）。下载所有和你的蛋白相关的 DNA 数据库条目到 GCG 中，然后进行 TFastA 搜索，将每个序列表达标签（EST）（如 DNA 序列）转换成蛋白，如果有一条 EST 和你蛋白家族中的任一个都不匹配，这时你就找到了一条新基因。

第五部分：多序列比对

GCG 有关 PileUp 程序的文档（>genhelp pileup）

PileUp 的功能

PileUp 采用渐进式的两两比对的方式将一组相关序列进行多序列比对，并且画聚类树图来帮助创建比对。

PileUp 描述

PileUp 运用 Feng-Doolittle 渐进式比对（J Mol Evol，1987，25，351～360）的一种简化形式来进行多序列比对。这种方法同 Higgins-Sharp（CABIOS，1989，5，151～153）描述的方法很相似。

多序列比对的步骤如下：先是两个最相似的序列之间两两比对，生成两比对序列的聚类，这个聚类再和剩余的最相关的序列或聚类比对，两个聚类的比对可以通过两条独立序列之间两两比对的简单延伸而得，最终比对结果的实现是进行一系列的渐进的两两比对，不断加入不太相似的序列或聚类，直到所有的序列都被加到最终的两两比对里。

在比对之前，序列先以相似度聚类生成一个系统树图，或一个反映序列聚类关系的树图。这个系统树图指导后续的两两比对的顺序。PileUp 可以将系统树图描绘出来，这样可以直观地看到最终比对结果产生过程中两两比对的次序。

一般而言，PileUp 可以比对多达 500 条序列，序列长度限制在 7000 个字符内（包括 PileUp 比对中插入的 gap 字符）。但是，如果要比对的序列比较长，PileUp 可比对的序列数目将会减少。

PileUp 举例

运行 PileUp 有两种简单的方法：

① 在命令行，敲入“>PileUp *.pep”，所有带 .pep 后缀的文件会加载到 PileUp 中。

② 若目录中有很多文件，你可以创建一个名为“anyname”的特殊文件，其中包括想要加载的具体蛋白或核酸序列。

>cat>anyname

a. pep

b. pep

c. pep

[∧Y]

>PileUp@anyname

这里有个输出的样本：

```
> pileup @x

PileUp creates a multiple sequence alignment from a group of related
sequences using progressive, pairwise alignments. It can also plot a
tree showing the clustering relationships used to create the alignment.

   1    humsyn1a.pep    289  aa
   2     ratsyn2.pep    291  aa
   3     humsyn3.pep    290  aa
   4     humsyn4.pep    298  aa
   5     humsyn5.pep    302  aa
   6     ratsyn6.pep    255  aa
   7     humsyn7.pep    261  aa
   8   humsyn16b.pep    324  aa

What is the gap creation penalty (* 12 *)  ?  8

What is the gap extension penalty (* 4 *)  ?  2

This program can display the clustering relationships graphically.
Do you want to:

    A) Plot to a FIGURE file called "pileup.figure"
    B) Plot graphics on HP7550 attached to /dev/tty15
    C) Suppress the plot

Please choose one (* A *):

The minimum density for a one-page plot is 6.7 sequences/100 platen units.
What density do you want (* 6.7 *) ?

What should I call the output file name (* x.msf *) ?

Determining pairwise similarity scores...

  1   x   2      3.15
  1   x   3      3.02
  1   x   4      2.09
         ...
  5   x   8      0.27
  6   x   7      0.15
  6   x   8      0.29
  7   x   8      0.79

Aligning...

  1      ..............-..
  2      ..............-.
         ..............-..
  3      ..............-.
         ..............-..
  4      ................-..
  5      ................-..
  6      ...............-..
  7      ............-..
```

如果你选择选项 A，程序会输出多序列比对结果，例如：

```
> cat x.msf
!!AA_MULTIPLE_ALIGNMENT 1.0
PileUp    of: @x

 Symbol  comparison   table: GenRunData:blosum62.cmp   CompCheck:  6430
                      GapWeight: 8
                GapLengthWeight: 2

 x.msf   MSF: 363    Type:  P  March 9, 1998  13:42  Check:  7155 ..

 Name:  humsyn1a              Len:   363  Check:  3616  Weight:   1.00
 Name:  ratsyn2               Len:   363  Check:  1439  Weight:   1.00
 Name:  humsyn3               Len:   363  Check:  7304  Weight:   1.00
 Name:  humsyn4               Len:   363  Check:  9241  Weight:   1.00
 Name:  humsyn7               Len:   363  Check:  8384  Weight:   1.00
 Name:  humsyn16b             Len:   363  Check:  2604  Weight:   1.00
 Name:  humsyn5               Len:   363  Check:  2749  Weight:   1.00
 Name:  ratsyn6               Len:   363  Check:  1818  Weight:   1.00

//
            1                                                    50
 humsyn1a   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~MKDRTQELR
  ratsyn2   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~MRDRLPDLT
  humsyn3   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~MKDRLEQLK
  humsyn4   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~MRDRTHELR
  humsyn7   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
humsyn16b   MATRRLTDAF LLLRNNSIQN RQLLAEQVSS HITSSPLHSR SIAAELDELA
  humsyn5   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~M SCRDRTQEFL
  ratsyn6   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~

            51                                                  100
 humsyn1a   TAK.DSDDDD .D.VAVTVDR DRFM.....D EFFEQVEEIR GFIDKIAENV
  ratsyn2   ACR.KSDDGD .NAVIITVEK DHFM.....D AFFHQVEEIR SSIARIAQHV
  humsyn3   AKQLTQDDDT .DAVEIAIDN TAFM.....D EFFSEIEETR LNIDKISEHV
  humsyn4   QGDDSSDEED KERVALVVHP GTARLGSPDE EFFHKVRTIR QTIVKLGNKV
  humsyn7   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~MSYTP GVGGDPTQLA QRISSNIQKI
humsyn16b   DDRMALVSGI SLDPEAAIGV TKRPPPKWVD GV....DEIQ YDVGRIKQKM
  humsyn5   SACKSLQTRQ N...GIQTNK PALRAVRQRS EFTLMAKRIG KDLSNTFAKL
  ratsyn6   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~MSM

            101                                                 150
 humsyn1a   EE..VK..RK HSAIL...AS PNPDEK..TK EELEELMSDI KKTANKVRSK
  ratsyn2   ED..VK..KN HSIIL...SA PNPEGK..IK EELEDLNKEI KKTANRIRGK
  humsyn3   EE..AK..KL YSIIL...SA PIPEPK..TK DDLEQLTTEI KKRANNVRNK
  humsyn4   QE..LE..KQ QVTIL...AT PLPEES..MK QELQNLRDEI KQLGREIRLQ
  humsyn7   TQ..CS..VE IQRTLNQLGT PQDSPE..LR QQLQQKQQYT NQLAKETDKY
humsyn16b   KE..SA..SL HDKHLNR.PT LDDSSE..EE HAIEITTQEI TQLFHRCQRA
  humsyn5   EK..LTILAK RKSLFDDKAV EIEELTYIIK QDINSLNKQI AQLQDFVRAK
  ratsyn6   EDPFFVVKGE VQKAVNTAQG LFQRWTELLQ GPSAATREEI DWTTNELRNN
```

如果你选择选项 B，程序会作树图（见下文）。

第六部分：系统树

何为系统树？

系统树是一个蛋白家族或其他家族（如基因家族）进化历史的图形表示（参见第 11 章）。系统树重建了一组蛋白的分支图谱，树的拓扑结构在一定程度上反映了蛋白沿时间的进化历史。

系统树的作用

系统树图将一组核酸或蛋白序列之间的关系进行图形化表示。假设你在研究克隆自小鼠 cDNA 文库的一个特定的激酶，系统树可以解答你以下几个问题：

① 你的激酶有没有高度同源的序列存在？例如，你敲除这个激酶的基因后有紧密相关的激酶吗？如果用药物改变了激酶的活性，有没其他的激酶活性也受到影响？

② 你应该如何命名之呢？

③ 它在进化中保守吗，还是序列变异得很快？这可让你一窥其功能。

④ 哪些氨基酸残基对于激酶的功能很关键？

系统树的做法

要做系统树，一开始要将你的蛋白或核酸序列进行 PileUp。PileUp 的输出结果可以是多序列比对的形式，带有 pileup. msf（multiple sequence format）的后缀。此外，你可以选择输出选项 B，那么 PileUp 输出的结果为简单的系统树。

我们在第 11 章中描述了如何做系统树。GCG 中一个关键的作树程序就是 distances，它将多序列比对中的两两序列比对做距离矩阵，另一个程序 Diverge 是在两个或多个比对序列蛋白编码区的每个位点上测量同义替换和非同义替换的数目，并将这些值以矩阵形式输出。GrowTree 重构系统树时可用距离矩阵或同义或非同义替换矩阵来实现。

第七部分：序列分析

Prime	设计寡核苷酸引物
Map	DNA 序列作图，显示限制酶
Map six	作图显示限制酶切割位点的六个残基
Motif x. pep	扫描 PROSITE 数据库，查找 x 蛋白中的 motif 和 domain

GCG 有关 Prime 程序文档（>genhelp prime）

功能

Prime 为模板 DNA 序列选择合适的寡核苷酸引物，引物用于 PCR（聚合酶链反应）或 DNA 测序。你可以设置 Prime 从整个模板中挑选引物，或者限制在文件列出的引物集合中挑选。

描述

Prime 分析模板 DNA 序列，然后挑选引物对进行 PCR 或者 DNA 测序。选择 PCR 引物对时，你可以选定模板序列上待扩增的区域。选择 DNA 测序引物，你可以指定模板上必须测序的位置。

GCG 有关 Map 的文档（>genhelp map）

功能

Map 用于 DNA 序列作图，显示 DNA 双链，并在序列上方标出限制酶切位点，下方标出翻译的蛋白。Map 也可用于氨基酸序列作图。

描述

Map 显示正在装配或正在详细分析的序列，Map 会让你选择哪种酶，并且该酶的限制性酶切位点会被标出，即在序列上面注上该酶的名字。如果你不回答这个问题，Map 会从所有商业酶中选出一个代表性的同裂酶。你可以选择将序列翻译成任何一种或所有六种可能的翻译框。你也可以选择只翻译开放阅读框。

运行 Map 后，在 Map 程序输出文件中，使用 ExtractPeptide 程序提取任何一个 DNA 翻译框翻译而成的多肽序列，来创建一个新序列文件。

Map 程序使用举例

我们输入一条 DNA 序列（human syntaxin 7），并且设置输出结果只显示酶切位点为 6 个残基的限制酶（158 个这样的限制酶，包括 *Eco*R Ⅰ，*Pvu* Ⅱ，*Bam*HI 等）。程序总共可以分析 500 多种限制酶的位点。注

意，你可以设置输出结果不显示预测蛋白，或显示所有 6 种可能的阅读框。

```
> map humsyn7.seq -six

Map maps a DNA sequence and displays both strands of the mapped sequence
with restriction enzyme cut points above the sequence and protein
translations below. Map can also create a peptide map of an amino acid
sequence.

                 Begin  (* 1 *) ?
               End  (*  1673 *) ?

Select the enzymes:  Type nothing or "*" to get all enzymes. Type "?"
for help on which enzymes are available and how to select them.

                                        Enzyme(* * *):

What protein translations do you want:

     a) frame 1   b) frame 2   c) frame 3
     d) frame 4   e) frame 5   f) frame 6

     t)hree forward frames   s)ix frames   o)pen frames only

     n)o protein translation   q) uit

Please select (capitalize for 3-letter) (* t *):

What should I call the output file (* humsyn7.map *) ?

Mapping ..................................

Writing .............. ..
MAP complete with:

  Sequence Length:      1,673
   Enzymes Chosen:        158
   Cutsites found:         94
         CPU time:      00.28

   Output file(s):    humsyn7.map

> cat humsyn7.map
(Linear)  (MinSite=6) MAP of: humsyn7.seq check: 612 from: 1 to: 1673

With 158 enzymes: *
                            October 29, 1998 12:18   ..

                                              MmeI
                                Bpu10I    TaqII  |
                  MspA1I       MspA1I|   BanII|  |
                   PvuII        PvuII|Bsp1286I|  |           PshAI
                       |            ||       ||  |               |
        CGCCACCGCCCATCAGCTGAGAATTGCAGCTGAGGGCTCCGGGGTAGGTGGGTGACGGCG
      1 ---------+---------+---------+---------+---------+---------+ 60
        GCGGTGGCGGGTAGTCGACTCTTAACGTCGACTCCCGAGGCCCCATCCACCCACTGCCGC
a       R  H  R  P  S  A  E  N  C  S  *  G  L  R  G  R  W  V  T  A  -
b        A  T  A  H  Q  L  R  I  A  A  E  G  S  G  V  G  G  *  R  R -
c         P  P  P  I  S  *  E  L  Q  L  R  A  P  G  *  V  G  D  G  G-
```

GCG 有关 Motifs 程序的文档（>genhelp motifs）

功能

Motifs 通过搜索蛋白模式来寻找 motif，这种模式的定义在 PROSITE 蛋白位点和模式字典中（参见第 8 章）。Motifs 可以显示它所查找到的每一篇关于 motif 的当前的文献摘要。

描述

Motifs 将你的蛋白序列和 PROSITE 字典中每一条序列模式进行比较来寻找蛋白序列 motif，Motifs 可以识别部分未匹配的序列模式（但不能有 gap）。目前，Motifs 只能寻找蛋白序列的模式。在 PROSITE 字典中，对每个 motif 都有一个简单的说明摘要。这些摘要显示在序列中找到的 motif 的旁边。

Motifs 程序使用举例

```
> motifs humsyn1a.pep

Motifs looks for sequence motifs by searching through proteins for the
patterns defined in the PROSITE Dictionary of Protein Sites and
Patterns. Motifs can display an abstract of the current literature on
each of the motifs it finds.

 What should I call the output file (* humsyn1a.motifs *) ?

        humsyn1a.pep len:       289 ...........................

             Total finds:         1
            Total length:       289
         Total sequences:         1
          CPU time (sec):     01.63
             Output file: "/user/pevsner/pevsner/syntaxin.dir/humsyn1a.motifs"

> cat humsyn1a.motifs
MOTIFS from: humsyn1a.pep

Mismatches: 0                    October 29, 1998 12:28 ..
       humsyn1a.pep     Check: 5551  Length: 289      ! LOCUS     HUMSYN1A     2088
bp    mRNA                PRI       23-NOV-1994
______________________________________________________________________

Epimorphin              Rx3(L,I,V,M)x2(L,I,V,M)Ex2(L,I,V,M)xE(L,I,V,M)x2(L,I,V,M)F
x2(L,I,V,M)x2(L,I,V,M,F)Vx2Q(G,Q)x2(L,I,V,M)(D,N,Q)x(L,I,V,M)(D,E)xN
                          Rx{3}(I)x{2}(L)Ex{2}(I)xE(L)x{2}(M)Fx{2}(M)x{2}(L)Vx{2}Q
(G)x(2)(I)(D)x(I)(E)xN
           198: SEIET   RHSEIIKLENSIRELHDMFMDMAMLVESQGEMIDRIEYN    VEHAV
*******************************
* Epimorphin family signature *
*******************************

The following proteins have been shown to be evolutionary related [1,2,3]:

 - Epimorphin, a mammalian mesenchymal protein which plays an essential role in
   epithelial morphogenesis.
 - Syntaxin A (also known as antigen HPC-1) and syntaxin B which are synaptic
   proteins which may be involved in docking of synaptic vesicles at presynaptic
   active zones.
 - Yeast PEP12 which is required for the transport of proteases to the vacuole.
 - Yeast SED5 which is required for the fusion of transport vesicles with the
   Golgi complex.
 - Yeast SSO1 and SSO2 which are required for vesicle fusion with the plasma
   membrane.

The above proteins share the following characteristics: a size ranging from
30 Kd to 40 Kd; a C-terminal extremity which is highly hydrophobic and is
probably involved in anchoring the protein to the membrane; a central, well
conserved region, which seems to be in a coiled-coil conformation.
```

[redacted]

The pattern specific for this family is based on the most conserved region of the coiled coil domain.

-Consensus pattern: R-x(3)-[LIVM]-x(2)-[LIVM]-E-x(2)-[LIVM]-x-E-[LIVM]-x(2)-
[LIVM]-F-x(2)-[LIVM]-x(2)-[LIVMF]-V-x(2)-Q-[GQ]-x(2)-
[LIVM]-[DNQ]-x-[LIVM]-[DE]-x-N

-Sequences known to belong to this class detected by the pattern: ALL.

-Other sequence(s) detected in SWISS-PROT: NONE.

-Last update: October 1993 / First entry.

术语表

本术语表结合了下面5个基于网络的术语表，每个条目都是根据它们来注释的：①美国国家生物技术信息中心（National Center for Biotechnology Information，NCBI）；②美国橡树岭国家实验室（Oak Ridge National Laboratory，ORNL）；③美国国立人类基因组研究所（National Human Genome Research Institute，NHGRI）的有声术语表；④SMART数据库（SMART database）；⑤来自蛋白质结构分类数据网站（Structural Classification of Proteins website，SCOP）的蛋白质折叠词汇（这些条目已被修改）。这5个术语表的网址分别是：

http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html；

http://www.ornl.gov/TechResources/Human_Genome/glossary/；

http://www.genome.gov/page.cfm?pageID=10002096；

http://smart.embl-heidelberg.de/help/smart_glossary.shtml；

http://scop.mrc-lmb.cam.ac.uk/scop/gloss.html。

其他的网络术语表在本术语表的末尾以表格的形式列出。

A

Additive genetic effects：加性遗传效应。不同位置上等位基因的联合效应等于它们的单独效应之和。(ORNL)

Adenine (A)：腺嘌呤。一种含氮碱基，是碱基对腺嘌呤-胸腺嘧啶（adenine-thymine，A-T）的一部分。参见 base pair。(ORNL)

Algorithm：算法。在计算机程序中包含的一种固定过程。(NCBI)

Alignment：序列比对。(a) 将两个或多个序列排在一起，以达到最大一致性的过程（对于氨基酸序列是比较它们的保守性），这样可以评估序列间的相似性和同源性。(NCBI)。(b) 将氨基酸在三维上重叠，从而对氨基酸的三级结构同系物进行预测的手段。SMART 中的比对大多基于已发表的观测结果（详细情况参见结构域注释），并且通过手工方式更新和编辑。(SMART)

All alpha：全α型。这一类型蛋白的结构域或共同核心部分的二级结构通常是3-，4-，5-，6-或多个α螺旋。(SCOP)

All beta：全β型。这一类型蛋白包括两个主要的折叠类型：夹层结构（sandwiches）和桶状结构（barrels）。夹层结构由两个互相扭曲而排成一列的β折叠组成。桶状结构由单个β折叠通过扭转和卷曲形成，大多数情况下β折叠的第一个分支链以氢键形式与最后一个分支链相连接。桶状结构两侧的分支链几乎是垂直的。这种具有垂直结构的β折叠在少数特殊的夹层结构中也存在。(SCOP)

Allele：等位基因。(a) 遗传位点上一对基因中的一个；遗传位点上的单个等位基因由亲本中的一个所遗传（例如，控制眼睛颜色位点上的等位基因可能产生蓝色或棕色眼睛）。(ORNL)。(b) 位于一对同源染色体的相同位置上控制某一性状的不同形态的基因。不同的等位基因产生诸如发色或血型等遗传特征的变化。在个体中，等位基因的某个形式（显性的）可以比其他形式（隐性的）表达得多。(NHGRI)

Allogeneic：同种异基因的。(ORNL)

Alternative splicing：可变剪接。基因外显子通过不同方式的结合产生不同的完全蛋白。(ORNL)

Amino acid：氨基酸。生物体中参与构成蛋白质的一类分子，有20种。蛋白质中的氨基酸序列和蛋白质的功能由遗传密码决定。(ORNL)

Amplification：扩增。对某种特定 DNA 片段拷贝数目增加的方法，有体内扩增和体外扩增两种。参

见 cloning。(ORNL)

Animal model：动物模式。参见模式生物(model organisms)。(ORNL)

Annotation：注释。对数据库中原始的 DNA 碱基序列添加相关的信息(比如编码的基因、氨基酸序列等)或其他的注解。参见 bioinformatics。(ORNL)

Anticipation：早先遗传。指某种遗传病在连续世代中，病情一代比一代严重的现象；例如，一个小孩可能比他的双亲发病时间要早、病情要严重；而他的双亲可能比他的祖父母发病更早。参见 additive genetic effects、complex trait。(ORNL)

Antisense：反义核酸。反义核酸序列与机体中产生的 mRNA 刚好相反；它结合在 mRNA 分子上可以组织蛋白质的形成。参见 transcription。(ORNL)

Apoptosis：细胞凋亡。细胞程序性死亡，是机体对受损的、不再需要的、无用的细胞的正常除去方法。

Array (of hairipins)：发夹序列。一类不能描述成束状或折叠叶的 α 螺旋。

Arrayed library：克隆矩阵。以二维形式排列在微滴定板上的单个初级的重组体克隆(宿主可以是噬菌体、质粒、YAC 或其他载体)。每个初级克隆能根据微滴定板的特性和板上克隆定位(行或列)而被确定。克隆阵列库有多方面的应用，包括筛选特殊基因或目的基因组区域等。参见 library、genomic library、gene chip technology。(ORNL)

Assembly：装配。将排好序的 DNA 片段插入染色体中的正确位置。(ORNL)

Autoradiography：放射自显影术，是用 X 光片显示放射性元素标记的分子及分子片段位置的技术；常用在使用凝胶将 DNA 片段分离之后，分析各个 DNA 片段的长度和数目。

Autosomal dominant：常染色体显性。位于一对同源常染色体上相同座位的等位基因有显性与隐性之分，所以在显性纯合与杂合状态下，表现出与显性基因有关的性状。后代在受精过程中获得该基因的概率为 50%。(ORNL)

Autosome：常染色体。和性别决定无关的染色体。人是双倍体，人类基因组含有 46 条染色体：其中 22 对是常染色体，一对是性染色体(XX 或 XY 染色体)。参见 sex chromosome。(ORNL)

B

Backcross：回交。将由两个亲本杂交所得的杂种子代再与两亲本中的一个进行杂交。在育种方面也指回交之后的异交。参见 model organisms。(ORNL)

Bacterial artificial chromosome(BAC)：细菌人造染色体。(a) 一种用于在大肠杆菌细胞(*Escherichia coli*)中克隆 DNA 片段的载体(其插入片段长约 100～300kb，平均长度约 150kb)。在大肠杆菌中存在天然质粒——F 因子。参见 cloning vector。(ORNL)。(b) 被克隆到细菌中的来自其他物种的长度为 100000～200000 个碱基对的大片段 DNA。外来的 DNA 一旦被克隆到宿主细菌中，就会产生大量拷贝。(NHGRI)

Bacteriophage：噬菌体。参见 phage。(ORNL)

Barrel：桶状结构。这种结构通常由 β 折叠的首末分支链之间的氢键所围成；它由两个整数来定义：β 折叠的分支链数 n 以及反映折叠中分支链交错程度的剪力参数 S。(SCOP)

Base：碱基。组成 DNA 和 RNA 的一类分子。参见 nucleotide、base pair、base sequence。(ORNL)

Base pair(bp)：碱基对。两个含氮碱基(A 和 T，或者 C 和 G)之间靠氢键结合在一起，形成一个碱基对。DNA 的两条链就是靠碱基对之间的氢键连接在一起，形成双螺旋结构。(ORNL)

Base sequence：碱基序列。DNA 分子中含氮碱基的排列顺序；它决定了由该 DNA 所编码的蛋白的结构。(ORNL)

Base sequence analysis：碱基序列分析。一种分析出 DNA 分子中碱基序列的方法(这种方法有时能够全自动化)。(ORNL)

Behavioral genetics：行为遗传学。研究可能影响行为的基因的科学。(ORNL)

Beta(β) sheet：β 折叠。有三种类型：反平行型(例如，相邻两个分支链的方向是反平行的)、平行型(指 β 折叠中所有分支链彼此平行的情形)和混合型(指在 β 折叠中，至少有一个分支链和相邻的一个分支链平行，并和相邻的另一个分支链反平行的情形)。(SCOP)

Bioinformatics：生物信息学。(a) 一门结合生物技术和信息技术从而揭示生物学中新原理的科学。(NCBI)。(b) 一门使用高级计算机技术来管理和分析生物学数据的科学。它在分析基因组研究数据时尤为重要。(ORNL)

Bioremediation：生物治疗。利用生物(如植物或微生物)去除一个区域的有毒物质。(ORNL)

Biotechnology：生物技术。一套由生物学基本实验而开发的技术，现在已用于生物学研究和产品开

发。生物技术也特指使用重组 DNA 工艺、细胞融合以及新的生物工艺的技术。

Birth defect：先天缺陷。由基因突变或出生前的非遗传因素而引起的出生时就存在的任何身体上或生化上的缺陷。

Bit score：二进制值。(a) 二进制值 S' 源于统计性质被数量化的打分系统中产生的原始比对分数 S。由于二进制值相对于打分系统已经被标准化，它们常用于比较不同搜索之间的比对分数。(NCBI)。(b) 在 HMMer 和 BLAST 中，比对分数常采用二进制值。将下面两种似然性相除：即将检测序列看作数据库中同类序列的似然性，与用一个随机模型代替该检测序列产生的似然性相除。对这两个似然性的相除的结果取以 2 为底数的对数，就可以得到二进制值。(SMART)

BLAST：(a) 基本局部相似性比对搜索工具。在序列数据库中快速查找与给定序列具有最优局部对准结果的序列的一种序列比对算法。初步搜索是对打分至少为 T、长度为 W 的词进行的。打分的过程是用一个替代矩阵对查询序列和该词作比较。然后词长可以试着向两端伸长以获得一个超过阈值 S 的打分。参数 T 反映了搜索的速度大小和敏感性。可以参见 BLAST 用户指南（Query 或 BLAST）和 BLAST 使用指导来获得更详细的信息。(NCBI)。(b) 一种确定不同物种（如人类、果蝇或线虫）之间的同源（相似）基因的计算机程序。(ORNL)

BLOSUM：模块替换矩阵。在替换矩阵中，每个位置的打分是在相关蛋白局部对比模块中观察到的替换的频率而获得的。每个矩阵被修改成一个特殊的进化距离。例如，在 BLOSUM62 矩阵中，是使用一致性不超过 62%的序列进行配对来获得打分值的。一致性大于 62%的序列在配对时用单个序列表示，以避免过于强调密切相关的家族成员。

Bundle：束状结构。在这组 α 螺旋中，每个螺旋的方向都大致绕同一个轴。如果有螺旋结构，那么每个螺旋与轴成正角时为左手螺旋；成负角时为右手螺旋。(SCOP)

C

Cancer：癌症。指正常细胞不受抑制地分裂与生长的疾病。癌症可以从它的原位向身体的其他部位扩散，并且可能是致命的。参见 hereditary cancer、sporadic cancer。(ORNL)

Candidate gene：候选基因。指位于染色体上并被推测与某种疾病有关的基因。参见 positional cloning，protein。(ORNL)

Capillary array：毛细管阵列。用来分离 DNA 测序片段的、充满凝胶的硅毛细管。由于毛细管的直径很小，可以使用比传统的凝胶板较高的电场，并且可以提供高速、高通量的分离。(ORNL)

Carcinogen：致癌物质。某种通过改变细胞中的 DNA 引起癌症的物质。参见 mutagen。(ORNL)

Carrier：载体。一个具有未表达的隐性性状的个体。(ORNL)

cDNA library：cDNA 库。是编码基因 DNA 序列的集合。在实验室这些序列是从 mRNA 序列产生的。参见 messenger RNA。(ORNL)

Cell：细胞。进行生命活动的生命体的基本单位。参见 genome、nucleus。(ORNL)

Centimorgan：厘摩。一种度量重组概率的单位。在生殖细胞形成的减数分裂过程中，常常会发生同源染色体之间的交叉现象，如果两个标记之间发生交叉的概率为 1%，那么它们之间的距离就定义为 1cM。对人类来说，1cM 大致相当于 1Mbp。参见 megabase。(ORNL)

Centromore：着丝点。在细胞的有丝分裂过程中，从细胞的两端发出纺锤丝，连接在染色体的着丝点上，将染色体拉向细胞的两级。(ORNL)

Chimera (plural chimaera)：嵌合体。一种含有不同基因型细胞或组织的生物。这可能是宿主生物的突变细胞，或不同生物或不同种类的细胞。(ORNL)

Chloroplast chromosome：叶绿体染色体。在植物光合作用的器官（叶绿体）而不是细胞核中所发现的环状 DNA，它有大部分遗传物质。

Chromosomal deletion：染色体缺失。部分染色体上的 DNA 的缺失。(ORNL)

Chromosomal inversion：染色体倒位。一个染色体上同时出现两处断裂，中间的片段倒转 180°重新连接起来，随之使这一片段上的基因排列的顺序颠倒。(ORNL)

Chromosome：染色体。细胞核中能够自我复制的部分，包含承载核苷酸序列遗传信息的 DNA 分子。原核生物中染色体的 DNA 呈环状，整个基因组都在一个染色体上。真核生物中一般包含多个染色体，每条染色体都由 DNA 和不同种类的蛋白质构成。(ORNL)

Chromosome painting：染色体作图。用荧光染料来标记染色体的部位。常用于诊断疑难疾病，如白

血病。(ORNL)

Chromosome region p：p 染色体区域，指染色体的短臂部分。(ORNL)

Chromosome region q：q 染色体区域，指染色体的长臂部分。(ORNL)

Clone：克隆。一种对诸如 DNA 片段（如基因或其他区域）、整个细胞或整个生物等生命物质进行的精确拷贝的过程。(ORNL)

Clone bank：克隆库。参见 genomic library。(ORNL)

Cloning：克隆化。使用专门的 DNA 技术来产生大量的单个基因或其他 DNA 片段的拷贝的过程，这样可以为进一步研究获得足够的材料。这个过程在人类基因组计划中特指对 DNA 的克隆。将克隆得到的 DNA 分子收集，可得到克隆文库。第二种克隆对细胞有丝分裂的过程进行研究，以获得整个细胞的拷贝。这些克隆的细胞组成了与原始细胞完全相同的细胞系。第三种克隆可获得完全相同的动物，如著名的苏格兰绵羊 Dolly。参见 cloning vector。(ORNL)

Cloning vector：克隆载体。通常采用从病毒、质粒或高等生物细胞中获取的 DNA 作为克隆载体，在载体上插入合适大小的外源 DNA 片段，并注意不能破坏载体的自我复制性质。将重组后的载体引入宿主细胞中，并在宿主细胞中大量繁殖。常见的载体有质粒、噬菌粒、酵母人工染色体。载体通常是包含不同来源的 DNA 片段的重组分子。(ORNL)

Closed, Partly Opened, and Opened：关闭构型，部分开放构型和开放构型。对 α 型结构，这些构型都是反映疏水核被 α 螺旋所遮蔽的程度的。开放构型与疏水核有至少大于一个螺旋的空间。(SCOP)

Code：密码。参见 genetic node。(ORNL)

Codominance：共显性。指一个遗传特性的两个不同等位基因都表达的现象。参见 autosomal dominant, recessive gene。(ORNL)

Codon：密码子。参见 genetic code。(ORNL)

Coisogenic or congenic：同源系。与某种生物几乎相同的系；它们只在单个位点上发生改变。(ORNL)

Comparative genomics：比较基因组学。将人类基因组与模式生物（如小鼠、果蝇以及大肠杆菌）进行比较研究。(ORNL)

Complementary DNA (cDNA)：互补 DNA。以信使 RNA 为模板合成的 DNA。

Complementary sequence：互补序列。能够与其他 DNA 片段根据碱基互补规则（A 与 T 配对，C 与 G 配对）形成双链结构的核苷酸序列。如 GTAC 的互补序列是 CATG。(ORNL)

Complex trait：复杂性状。指基因组分不严格遵守孟德尔遗传定律的性状。可能包括两个和更多基因之间的相互作用，或者是基因和环境之间的相互作用。参见 Mendelian inheritance、additive genetic effects。(ORNL)

Computational biology：计算生物学。参见 bioinformatics。(ORNL)

Confidentiality：机密性。遗传学上，在没有提供者的允许时不能获得遗传学的资料和信息。通过检测对材料的获得程度可以看出材料的机密性。(ORNL)

Congenital：先天的。出生时由于遗传因素或非遗传因素所表现的形状。参见 birth defect。(ORNL)

Conservation：保守。指氨基酸或 DNA（普遍性较小）序列某个特殊位置上的改变，并不影响原始序列的物理化学性质。(NCBI)

Conserved sequence：保守序列。指 DNA 分子中的一个核苷酸片段或者蛋白质中的氨基酸片段，它们在进化过程中基本保持不变。(ORNL)

Contig：邻接片段。一组在染色体上有重叠区域的 DNA 片段的克隆。(ORNL)

Contig map：邻接图谱。邻接图谱描述覆盖了整个染色体的小片段的顺序关系，这些小片段相互邻接，可以通过两个片段有重叠部分推断出两者相互邻接。(ORNL)

Cosmid：黏端质粒。人工构造的含有 Lambda 抗菌素的 *cos* 基因的克隆载体。黏端质粒能够引入到 Lambda 抗菌素微粒中，然后注入到大肠杆菌中去，这样我们就可以将长达 45kb 的 DNA 片段引入到宿主细菌的质粒载体中。(ORNL)

Crossing over：交叉。在减数分裂时，来自父本的染色体和来自母本的染色体有时会发生断裂，然后交换断裂部分重新组合成新的染色体，这种交叉常常会导致等位基因的交换。参见 recombination。(ORNL)

Crossover：交叉。蛋白质结构核心的相反两端的二级结构之间的连接。这种交叉会穿过结构域的表面。(SCOP)

Cytogenetics：细胞遗传学。对染色体的物理性质进行研究的科学。参见 karyotype。(ORNL)

Cytological band：细胞带。染色体上与周围部位染色不同的区域。参见 cytological map。(ORNL)

Cytological map：细胞学图谱。一种根据染色体突变得到的细胞学上的发现而绘制的染色体图谱。(ORNL)

Cytoplasmic trait：细胞质的遗传特性。一种在细胞核外的叶绿体和线粒体中发现基因的遗传特性。结果使后代仅从单个亲本中就可获得某些遗传物质。(ORNL)

Cytoplasmic (uniparental) inheritance：细胞质遗传（单亲遗传）。参见 cytoplasmic trait。(ORNL)

Cytosine (C)：胞嘧啶。一种含氮碱基。在 DNA 中是碱基对 GC（鸟嘌呤-胞嘧啶）的一部分。参见 base pair，nucleotide。(ORNL)

D

Data warehouse：数据仓库。一种数据库、数据表格以及面向单个主题的数据集合。(ORNL)

Deletion：缺失。染色体上部分 DNA 丢失的现象，可能导致疾病或畸形。参见 chromosome，mutation。(ORNL)

Deletion map：缺失图谱。对用于定义突变的特殊染色体的一种描述，这种突变特异地筛除了基因组中的某些区域，作为"生物化学路标"或特殊区域的标记。(ORNL)

Deoxyribonucleotide：脱氧核糖核苷酸。参见 nucleotide。(ORNL)

Deoxyribose：脱氧核糖。一种糖，DNA（脱氧核糖核酸）的组分之一。(ORNL)

Diploid：双倍体。一整套包含成对染色体的遗传物质，一条来自父本，一条来自母本。大多数动物的细胞（配子细胞除外）都有包含双倍体的染色体。人类的双倍体基因组含有 46 条染色体。参见 haploid。(ORNL)

Directed evolution：定向进化。在实验室中对分离的分子或探针用突变的方法加以改造，以确定改变后的分子对新环境的适应。(ORNL)

Directed mutagenesis：定点诱变。改变 DNA 的一个特定碱基，并将该 DNA 重新插入生物体中，对改变后的影响进行研究。(ORNL)

Directed sequencing：定向测序。对邻近染色体上的 DNA 进行连续的测序。(ORNL)

Disease-associated genes：与疾病相关的基因。指携带与疾病相关的特定 DNA 序列的等位基因。(ORNL)

DNA (deoxyribonucleic acid)：DNA（脱氧核糖核酸）。编码遗传信息的大分子。DNA 是一种双链结构，两条链之间通过碱基对之间的氢键相互连接。DNA 中包含的四种碱基是腺嘌呤（A）、鸟嘌呤（G）、胞嘧啶（C）和胸腺嘧啶（T）。在自然界中，只有 A 与 T 之间、G 与 C 之间才能形成碱基对；因此我们能够通过一条链的顺序推断出另一条链的顺序。(ORNL)

DNA bank：基因库。一种用以存储从血液样本或其他人类组织中获得的 DNA 的服务。(ORNL)

DNA probe：DNA 探针。参见 probe。(ORNL)

DNA repair genes：DNA 修补基因。编码纠正 DNA 序列中错误的蛋白的基因。(ORNL)

DNA replication：DNA 复制。以现有 DNA 的一条链为模板合成一条新的链。在人类和其他真核生物细胞中，DNA 的复制在细胞核中进行。(ORNL)

DNA sequence：DNA 序列。DNA 片段、基因、染色体或整个基因组中的碱基排列顺序。参见 base sequence analysis。(ORNL)

Domain：结构域。(a) 蛋白质在折叠时与其他部分相独立的一个不连续的部分，它有自己独特的功能。(NCBI)。(b) 蛋白质中一个有着特定功能的独立单元。多个结构域共同决定蛋白质的功能。(ORNL)。(c) 有不同二级结构和疏水核的保守的结构实体。在富含二硫化物的小结构域和结合 Zn^{2+} 或 Cu^{2+} 的结构域中，其疏水核可能由胱氨酸和金属离子来提供。有相同功能的同源结构域通常在序列上有相似性。(SMART)

Domain composition：结构域组成。有相同结构域的蛋白质对于待检验的每个结构域至少有一个拷贝。(SMART)

Domain organization：结构域的组织。含有检测序列的所有结构域的蛋白质（附加的结构域也包含在内）。(SMART)

Dominant：显性的。即使只有一个拷贝存在时也总表达的等位基因。参见 gene、genome。(ORNL)

Double helix：双螺旋。DNA 的两条链互相缠绕在一起，形成一种双螺旋式的阶梯结构；两条链上的互补核苷酸以氢键连接在一起。(ORNL)

Draft sequence：初稿序列。由人类基因组计划产生的、并未完成的序列，它绘制了人类基因约 95% 的草图。初稿序列库几乎都是长约 10000bp 的片段，它们在染色体上的大概位置是已知的。(ORNL)

DUST：一个低复杂性区段过滤程序。(NCBI)

E

***E* value**：E 值。(a) 期望值。在一个数据库中所搜索到的打分值等于或大于 S 的不同比对的个数。E 值越低，表明该打分值的显著性越好。(NCBI)。(b) E 值表示的是所期望的打分值大于或等于 X 的序列的个数。用户提供的序列和数据库中的序列比对，会由任何算法产生一个打分值 X，E 值与这个打分值 X 相关，并且反映了从一个相同规模的随机序列数据库中搜索到的打分值接近或大于 X 的比对的个数。从 2.0 版本起，E 值都是用隐马尔可夫模型计算得到的，这与以前的算法相比，正确性有很大提高。(SMART)

Electrophoresis：电泳技术。分离大分子的一种方法，能够从一堆混杂在一起的 DNA 或者蛋白质中依据各个片段的大小将它们分开。一般在介质两端加电压。介质一端设有小槽，槽内放有待分离的大分子溶液，在电场的作用下，大分子会从一端向另一端运动，但是由于自身所带电量的大小或分子量的不同，它们的泳动速度是不同的，因此我们可以根据它们的位置将它们分离开来。分离蛋白质和核酸时常用的介质有琼脂糖和聚丙烯酰胺凝胶。(ORNL)

Electroporation：电穿孔技术。它利用高压电脉冲引起细胞膜双分子层产生通透性，以易于引入新的 DNA 分子；它通常用于重组 DNA 技术。参见 transfection。(ORNL)

Embryonic stem (ES) cells：胚胎干细胞。一种能无限复制的胚胎细胞，它能转化成其他类型的细胞，并且可以持续不断地产生新细胞。(ORNL)

Endonuclease：核酸内切酶。参见 restriction enzyme。(ORNL)

Enzyme：酶。一种特殊的具有催化作用的蛋白质，它能够加快生化反应的速度，但是不改变反应的方向和产物。(ORNL)

Epistasis：上位性。一对基因影响或组织了另一对非等位基因的效应，这种非等位基因间的相互作用方式称为上位性。(ORNL)

Escherichia coli：大肠埃希菌，大肠杆菌。细菌的一种。遗传学家对大肠杆菌研究得比较透彻，大肠杆菌的基因组比较小，通常没有致病性，易于培养。

Eugenics：优生学。借助人工选择来改善某个种类素质的学科；通常指改变人类的素质。(ORNL)

Eukaryote：真核生物。细胞或生物自身有细胞膜包被，有结构独立的细胞核，以及发育完全的细胞器。除了病毒、细菌和蓝藻绿藻外，绝大多数生物都是真核生物。参见 prokaryote，chromosome。(ORNL)

Evolutionarily conserved：保守进化的。参见 conserved sequence。(ORNL)

Exogenous DNA：外源 DNA。被引入某种生物体的、来源于其他生物体的 DNA。(ORNL)

Exon：外显子。基因中编码蛋白质的 DNA 序列。参见 intron。(ORNL)

Exonuclease：外切核酸酶。一种能连续将线性核酸酶自由端的核苷酸切割下来的酶。

Expressed gene：表达基因。参见 gene expression。(ORNL)

Expressed sequence tag (EST)：表达序列标记。一种短的 DNA 片段，是 cDNA 分子的一部分，可以用来鉴定基因，通常用于基因定位和基因图谱中。参见 cDNA，sequence-tagged site。(ORNL)

F

FASTA：(a) FASTA 是第一个被广泛使用的数据库相似性搜索算法。这个程序通过扫描序列中“词”的小配对，从而寻找最优局部比对。首先计算有多个序列的片段的分值（记为 init1）；接着这些分值加在一起产生“initn”值；最后输出包含间隔的最佳比对（记为 opt）。搜索的敏感性和速度由“词”的长度——“k-tup”变量所控制，并且二者负相关。(NCBI)。(b) 核苷酸或蛋白质序列的一种输出格式。

Filial generation (F_1, F_2)：子代。在繁殖过程中每个世代所产生的后代，记为 F_1，F_2 等。(ORNL)

Filtering：过滤，也叫掩蔽（masking），指对那些经常产生乱真的高分数的核苷酸或氨基酸序列区域进行隐藏的过程。

Fingerprinting：“指纹”识别。在遗传学上，对一个人 DNA 上多个特定等位基因的确定，可以作为确定该人身份的惟一标志。

Finished DNA sequence：完成序列。高质量、低错误率、无间隙的人类基因组 DNA 序列。达到 2003 年人类基因组计划（HGP）的最终目标，需要附加的排序来关闭序列间的间隙、减少模糊度，并且对于每 10000 个碱基只允许有一个误差（这是 HGP 完成测序的执行标准）。参见 sequencing，

draft sequence。(ORNL)

Flow cytometry：流式细胞术。一种生物材料的分析法，其主要原理是通过物质对激光的吸收特性或其荧光特性来检测该物质。当含细胞或亚细胞的片段通过一窄孔时，给予一束激光，会产生吸收谱或荧光谱。自动分拣装置依据每份样品的荧光发射特点对连续液滴所构成的样品进行分离。(ORNL)

Flow karyotyping：流式核型分析。利用流式细胞仪，在DNA的基础上分析和分离染色体。(ORNL)

Fluorescent in situ hybridization (FISH)：荧光标记原位杂交。一种物理图谱绘制方法，使用荧光素标记探针，以检测探针和分裂中期的染色体或分裂间期的染色质的杂交。(ORNL)

Folded leaf：折叠叶。环绕单个疏水核而形成的一层α螺旋，但并不是简单的桶状结构。(ORNL)

Forensics：法医鉴定。使用DNA来进行鉴定。使用DNA的例子有，在子女抚养案件中确定亲子关系，在犯罪现场确定嫌疑犯身份，以及确定事故受害者身份等。(ORNL)

Fraternal twin：异卵双生。由于两个卵子分别受精而同时出生的兄弟姐妹。就像同胞兄弟一样，他们彼此之间有相同的遗传关系。参见 identical twin。(ORNL)

Full gene sequence：全基因序列。指基因中已完成的碱基测序。这个顺序取决于基因所编码的蛋白。(ORNL)

Functional genomics：功能基因组学。一门研究基因、基因所编码的蛋白，以及该蛋白在身体的生化过程中所起的作用的学科。(ORNL)

G

Gamete：配子。成熟的雄性或雌性生殖细胞（精子或卵子），只有单倍体的染色体（对于人类来说，有23条染色体）。(ORNL)

Gap：空位。(a) 在两条序列比对过程中需要在检测序列或目标序列中引入空位，以表示插入或删除。为了避免在比对时出现太多的空位，可以在引入空位的同时，从比对的打分值中减去一个固定值（空位值）。在多余的核苷酸或氨基酸周围引入空位时，也要对比对的打分值进行罚分。(NCBI)。(b) 在两个序列比对时，空位代表缺失。为了减少大量连续空位区域的形成，需要在比对算法中加入空位罚分。空位也代表比对时由于蛋白质结构中突出的环或β凸出而形成的插入。(SMART)

GC-rich area：GC富含区域。许多DNA序列含有长的G、C重复片段，这些片段通常是富含基因的区域。(ORNL)

Gel electrophoresis：凝胶电泳。参见 electrophoresis。(ORNL)

Gene：基因。遗传的基本结构和功能单位。基因是特定染色体上特定位置的一段核苷酸片段，它能够编码具有特定功能的产物（如蛋白质或RNA分子）。参见 gene expression。(ORNL)

Gene amplification：基因扩增。一种增加特殊DNA片段拷贝数量的方法；是肿瘤细胞的一个特点。参见 gene、oncogene。(ORNL)

Gene chip technology：基因芯片技术。从大量的基因中形成cDNA微阵列的技术；用于监控和检测芯片上每个基因表达量的变化。(ORNL)

Gene expression：基因表达。基因编码的信息转化为细胞结构并在细胞中行使功能的过程。表达的基因包括转录成信使RNA接着翻译成蛋白质的基因，以及转录成RNA但不翻译成蛋白质（如转移RNA和核糖体RNA）的基因。(ORNL)

Gene family：基因家族。几组关系紧密、表达产物相似的基因。(ORNL)

Gene library：基因文库。参见 genomic library。(ORNL)

Gene mapping：基因图谱。在一个DNA分子（染色体或质粒）上决定基因的相对位置及其相互间距离的图谱。包括遗传图谱或物理图谱。(ORNL)

Gene pool：基因库。特定时间内一个特定物种的所有基因，包括它们的变种。参见 allele，gene、polymorphism。(ORNL)

Gene prediction：基因预测。用计算机程序对可能的基因所作的预测，它是基于DNA片段与已知基因序列的匹配程度的。(ORNL)

Gene product：基因产物。生物化学物质，可以是RNA也可以是蛋白质，它们是基因表达的产物。基因产物的数量能衡量这个基因的活性；而异常的产物数量与等位基因引起的疾病有关。(ORNL)

Gene testing：基因测试。参见 genetic testing、genetic screening。(ORNL)

Gene therapy：基因治疗。用于治疗遗传疾病的正在发展中的技术。医学程序包括基因替代、基因改造或用健康的基因补充无功能、错功能的基因等。参见 gene、inherit、somatic cell gene therapy、germ line gene therapy。(ORNL)

Genetic code：遗传密码。信使 RNA 上每三个一组的核苷酸序列，决定了合成蛋白质时肽链上的一个氨基酸。基因上的 DNA 序列控制形成信使 RNA 上的核苷酸序列，进而决定了蛋白质肽链上的氨基酸序列。(ORNL)

Genetic counseling：遗传咨询。对患有遗传疾病或有得这种疾病风险的个人和家庭进行短期的教育咨询。遗传咨询提供给患者关于他们情况的信息，并帮助他们做出明智的决定。(ORNL)

Genetic discrimination：基因歧视。指对已经有或可能有遗传障碍的人怀有偏见。(ORNL)

Genetic engineering：遗传工程。改变细胞或生物体的遗传物质，从而使他们产生新的物质或行使新的功能。(ORNL)

Genetic engineering technology：遗传工程技术。参见 recombinant DNA technology。(ORNL)

Genetic illness：遗传疾病。指由于一个或多个有害等位基因的遗传而造成的疾病、残障或其他缺陷。(ORNL)

Genetic informatics：遗传信息学。参见 bioinformatics。(ORNL)

Genetic map：遗传图谱。参见 linkage map。(ORNL)

Genetic marker：遗传标记。遗传特性能被跟踪的基因或其他可确认特性的 DNA 片段。(ORNL)

Genetic material：遗传物质。参见 genome。(ORNL)

Genetic mosaic：镶嵌型。指在同一个生物体中，不同的细胞含有不同的基因序列的现象。这可能是由发育过程中的突变或发育早期胚胎的融合造成的。(ORNL)

Genetic polymorphism：遗传多态性。指个体、类群、种群之间 DNA 序列（如蓝色眼睛和棕色眼睛的基因）的差异。(ORNL)

Genetic predisposition：遗传易感性。某种遗传疾病的易感性。在发育中可能会、也可能不会引起疾病。(ORNL)

Genetic screening：遗传筛选。测试一定数量的人以确定其中高风险患有或遗传特殊遗传病的个人。(ORNL)

Genetic testing：遗传测定。对个体的遗传物质进行分析，以此来确定该个体在特定健康条件下的体质或者以此来诊断遗传疾病。(ORNL)

Genetics：遗传学。研究物质的遗传特性模式的科学。(ORNL)

Gene transfer：基因转移。外源性的 DNA 由载体（如已被修饰的病毒）转入某受体菌细胞的过程，常用于基因治疗。参见 mutation、gene therapy、vector。(ORNL)

Genome：基因组。一种生物所有染色体上的遗传物质称为基因组，基因组的大小常常采用碱基对的数目来表示。(ORNL)

Genome project：基因组计划。基因组计划通过努力研究和发展技术，从而对人类或某些模式生物的基因组进行绘制图谱或测序。(ORNL)

Genomic library：基因组文库。代表整个生物体基因组的随机产生的重叠 DNA 片段的克隆的总和。(ORNL)

Genomics：基因组学。对基因和它们的功能进行研究的科学。(ORNL)

Genomic sequence：基因组序列。参见 DNA。(ORNL)

Genotype：基因型。根据生物体的物理性质（表现型）所能区分的基因组成。(ORNL)

Germ cell：生殖细胞。精子和卵子，以及它们的前体。生殖细胞是单倍体，只有一套染色体（共有 23 条），而其他的细胞是双倍体（共有 46 条染色体）。(ORNL)

Germ line：种系。一套遗传信息的延续传代过程。参见 inherit。(ORNL)

Germ line gene therapy：生殖细胞基因治疗。将外源正常基因转入精子、卵子或受精卵，从而改变生殖细胞中的某个病态遗传物质，防止后代患这种遗传病的基因工程。它可用来缓解与遗传病有关的疾病。参见 genomics、somatic cell gene therapy。(ORNL)

Germ line genetic mutation：生殖细胞基因突变。参见 mutation。(ORNL)

Global alignment：整体联配。对两个核苷酸或蛋白质序列的全长所进行的比对。(NCBI)

Greek key：希腊钥匙结构。一些 β 折叠所组成的拓扑结构，其中一些中间的分支链会穿过桶状结构的末端或穿过由 β 折叠所形成的三明治结构。(SCOP)

Guanine (G)：鸟嘌呤。一种含氮碱基。在 DNA 中是碱基对 GC（鸟嘌呤-胞嘧啶）的一部分。参见 base pair、nucleotide。(ORNL)

H

H：相对熵值。目标残基和底物残基频率的相对熵

记做 H。H 可以衡量某个位置（这个位置可以通过概率来区分比对）上由于偶然因素而得到的平均信息（用字节表示）。H 值越高，短的比对就越可以通过概率来区分；H 值越低，需要的比对长度越长。(NCBI)

Haploid：单倍体。单倍体细胞中只有一套染色体（是体细胞中染色体数目的一半），如动物的精子和卵子、植物的卵细胞和花粉都是单倍体细胞。人类的生殖细胞有 23 条染色体。参见 diploid。(ORNL)

Haplotype：单体型。用来表示一条染色体上一系列邻近位置上共同基因型的方法。(ORNL)

Hemizygous：半合子的。半合子只具有决定某一特征的一对基因中的一个。如在人类中，男性的 Y 染色体是半合子。(ORNL)

Hereditary cancer：遗传性癌。与接受亲代突变的基因有关的癌症类型。(ORNL)

Heterozygosity：杂合性，异型接合性。在某一个或一些位点上关于某一特征具有不同的等位基因的状态。(ORNL)

Heterozygote：杂合子。参见 heterozygosity。(ORNL)

highly conserved sequence：高度保守序列。在几种不同类型的生物体中非常相似的 DNA 序列。参见 gene、mutation。(ORNL)

high-throughput sequencing：高通量测序。一种快速测定 DNA 中碱基顺序的方法。参见 sequencing。(ORNL)

Homeobox：同源（异型）框。指核苷的简短延伸，这种核苷的碱基序列实际上在所有基因中都是相同的。这种现象在许多生物体（如从果蝇到人类）中都能发现。在果蝇中，同源（异型）框能确定特定基因在发育过程中表达的时间。(ORNL)

Homolog：同系物。双倍体生物的一对染色体中的一个，或者是两个或更多种类中起源和功能相同的基因。(ORNL)

Homologous chromosome：同源染色体。有着相同的线性基因序列的一对染色体，分别来自父本和母本。(ORNL)

Homologous recombination：同源重组。同源染色体之间所进行的 DNA 片段的交换。(ORNL)

Homology：同源性。(a) 由共同的祖先所遗传得到的相似性。(NCBI)。(b) 在同一物种不同个体间或不同物种间蛋白质序列或 DNA 的相似性。(ORNL)。(c) 由于基因复制而从共同祖先得到的进化遗传。(SMART)

Homozygote：纯合子。在特殊的基因上携带两个相同等位基因（从每个亲代获得一个）的个体。(ORNL)

Homozygous：纯合的。参见 homozygote。(ORNL)

HSP：高打分值片段。在一个给定的搜索中，没有空位的局部比对能得到最高的比对打分值。(NCBI)

Human gene therapy：基因治疗。参见 gene therapy。(ORNL)

Human Genome Initiative：人类基因组行动计划。是自 1986 年起美国能源部（DOE）领导的几个项目的总称。包括：①建立某个已知染色体位置的 DNA 片段的顺序；②开发分析基因图谱和对 DNA 测序的算法；③开发 DNA 检测和分析的新技术和新设备。能源部行动计划现在的名称是人类基因组计划，而整个美国有关的工作则称为人类基因组项目，由美国能源部和国立卫生研究院共同领导。

Human Genome Project (HGP)：人类基因组项目，以前的名称是人类基因组行动计划。参见 Human Genome Initiative。(ORNL)

Hybrid：杂种。基因相异的双亲所繁衍的后代。参见 heterozygote。(ORNL)

Hybridization：杂交。两段互补的 DNA 单链，或者一段 DNA 单链和一段 RNA 依照碱基互补规则形成一条双链的过程。(ORNL)

I

Identical twin：同卵双生。指由单个受精卵所分裂产生的双胞胎，他们有相同的基因型。参见 fraternal twin。(ORNL)

Identity：一致性。两个（核苷酸或氨基酸）序列比对时不变部分的长度。(NCBI)

Immunotherapy：免疫疗法。利用免疫系统来治疗疾病的一种概念，例如，培育疫苗。免疫治疗也可认为是对免疫系统导致的疾病的治疗。参见 cancer。(ORNL)

Imprinting：印记。一种疾病的表型依赖于父母双方的疾病基因的现象。例如，Prader-Willi 综合征和 Angelman 综合征在遗传时，15 号染色体的相同部分是丢失的。当来自父本的 15 号染色体的补体丢失时，小孩患 Prader-Willi 综合征；但是来自母本的 15 号染色体的补体丢失时，小孩患 *Angelman* 综合征。(ORNL)

Independent assortment：独立分配。不同的相对

性状的遗传因子在减数分裂过程中，这一对因子与另一对因子的分离和组合互不干扰，各自独立分配到配子中去。参见 linkage。(ORNL)

Informatics：信息学。参见 bioinformatics。(ORNL)

Informed consent：知情同意。指向受试者告知一项试验的各个方面的情况后，受试者自愿确认其同意参加该项临床试验的过程。(ORNL)

Inherit：遗传。指遗传学上通过生命活动而从亲本获得遗传物质。(ORNL)

Inherited：遗传的。参见 inherit。(ORNL)

Insertion：插入。一种染色体畸变，由于一段 DNA 序列插入某个基因中，从而破坏该基因的正常结构和功能。参见 chromosome、DNA、gene、mutation。(ORNL)

Insertional mutation：插入突变。参见 insertion。(ORNL)

In situ hybridization：原位杂交。使用 DNA 或者 RNA 探针来检测与其互补的另一条链在细菌或其他真核细胞中的位置。(ORNL)

Intellectual property rights：知识产权。指专利权、版权以及商标专用权。参见 patent。(ORNL)

Interference：干扰。指一个交叉事件的发生，会使另一个交叉事件发生的可能性降低的现象。也称为正干扰。而负干扰是增加了第二个交叉事件发生的可能性。参见 crossing over。(ORNL)

Interphase：分裂间期。指在细胞周期中，DNA 在细胞核内复制的时期。它发生在有丝分裂之后。(ORNL)

Intracellular domains：细胞内区域。指在细胞质的蛋白质中普遍存在的结构域家族。(SMART)

Intron：内含子。断裂基因中外显子的间插序列，可参与前体 RNA 的转录，但其转录的 RNA 序列在转录后的加工中被切除，不包含于成熟的 RNA 分子中。参见 exon。(ORNL)

In vitro：体外。在活的生物体之外（如在实验室）所进行的研究。(ORNL)

In vivo：体内。在活的生物体之内进行的研究。(ORNL)

Isoenzyme：同工酶。功能相同、但氨基酸序列却不同的酶的不同化学形式中的一种。这两种酶可能会产生不同的反应速率。(ORNL)

J

Jelly roll："春卷"样空间结构。希腊钥匙（Greek-key）拓扑结构的一种变形，其三明治结构或桶状结构折叠的两端被两个中间分支链所连接。参见 Greek key。(SCOP)

Junk DNA：垃圾 DNA。不编码基因的 DNA 片段；大部分基因中所含的垃圾 DNA 可能有调节或其他的功能，也称为非编码 DNA。(ORNL)

K

K：*K* 值。用来计算 BLAST 程序中打分函数的一个统计参数；它可以看作搜索空间大小的一个自然衡量尺度。*K* 值通常用于将原始比对值 S 转换为二进制值 S'。(NCBI)

Karyotype：染色体组型。一组以标准模式排列的染色体显微摄影照片，显示了每个染色体的数量、大小、形状；并用以低分辨率的物理图谱显示染色体畸变和疾病的关联。(ORNL)

Kilobase (kb)：DNA 片段的长度单位，相当于 1000 个核苷。(ORNL)

Knock-out：剔除。特殊基因的灭活。剔除通常在实验室的生物体中进行，从而研究基因的功能。(ORNL)

L

Lambda：λ 值。用来计算 BLAST 程序中打分函数的一个统计参数；它可以看作打分系统的一个自然衡量尺度。λ 值通常用于将原始比对值 S 转换为二进制值 S'。(NCBI)

Library：库。一些克隆（如从特定生物体中克隆得到的 DNA）的无顺序的汇总；这些克隆之间的关系可以通过物理图谱来建立。参见 genomic library、arrayed library。(ORNL)

Linkage：连锁。染色体上两个或多个标记（如基因、限制性片段长度多态性标记）的接近度；标记越紧密，在 DNA 修复或复制过程中（原核细胞中的二分裂，真核细胞中的有丝分裂或减数分裂）它们分离的可能性就越低，因此它们在一起被遗传的可能性就越高。(ORNL)

Linkage disequilibrium：连锁不平衡。指某些遗传标记的基因重组发生在物种中的频率，高于或低于从其距离推测的值。这说明两对等位基因在 DNA 链上位置是紧密的。参见 Mendelian inheritance。(ORNL)

Linkage map：连锁图谱。描述染色体上遗传位点相对位置的图谱，是决定遗传位点同时遗传的频率

的重要基础。以厘摩（cM）作为距离测量单位。（ORNL）

Local alignment：局部联配。对两个核苷酸或蛋白质序列的一部分所进行的比对。（NCBI）

Localization：局部化。指在不同的细胞空间（细胞质、细胞外的空间、细胞核内，以及与膜相连的区域）中所表现的结构域的个数；在 SwissProt 数据库中它在注释中标出。（SMART）

Localize：基因定位。是对基因或其他标记在染色体上所在位置进行的测定。（ORNL）

Locus（plural loci）：基因座位。特殊基因或其他标记在染色体上的位置，也指 DNA 所在的位置。基因座位有时候可以用来限制 DNA 表达的区域。参见 gene expression。（ORNL）

Long-range restriction mapping：远距离限制性作图。限制性酶是一种在精确位点切割 DNA 的酶。限制性图谱显示了限制性酶的切割位点在染色体上的位置。它们在染色体上常用作生化“路标”或特定区域的标志。这张图谱详细显示了 DNA 被特定的限制性酶所切割的位置。（ORNL）

Low-complexity region（LCR）：低复杂性区域。指组分（包括均聚物、短周期重复片段）区域和有许多单个或多个残基的区域。SEG 程序用来筛选或过滤氨基酸序列中的低复杂性区域。DUST 程序用来筛选或过滤核苷酸序列中的低复杂性区域。（NCBI）

M

Macrorestriction map：限制性图谱。描述染色体上限制性内切酶切割位点之间距离和顺序的图谱。（ORNL）

Mapping：作图。参见 gene mapping、linkage map、physical map。（ORNL）

Mapping population：定位族群。指在构建一个遗传图谱时所用到的一组相关的生物。（ORNL）

Marker：标记。参见 genetic marker。（ORNL）

Masking：掩蔽。也叫过滤（filtering），指为了提高对序列相似性搜索时的敏感性，而从序列中移除重复的或低复杂性区域的过程。（NCBI）

Mass spectrometer：质谱仪。一种通过质量和电荷来确定化合物的装置。（ORNL）

Meander：回纹波形。一种简单的 β 折叠结构，其中任意两个连续的分支链是相邻和反平行的。（SCOP）

Megabase（Mb）：兆碱基。DNA 片段长度单位，相当于 100 万个核苷，大约等于 1cM。参见 centimorgan。（ORNL）

Meiosis：减数分裂。在倍数染色体起源的性细胞中发生的连续两次细胞分裂的过程。减数分裂的结果是产生 4 个子细胞，而不是 2 个，每个子细胞中仅带有一组染色体。参见 mitosis。（ORNL）

Mendelian inheritance：孟德尔遗传规律。基因所表现出的方式和特征是由父母遗传给子女的。它是以孟德尔名字命名的，他首先研究和提出了基因的存在，并提出了这个遗传规律。参见 autosomal dominant、recessive gene、sex linked。（ORNL）

Messenger RNA（mRNA）：信使 mRNA。蛋白质合成的模板。参见 genetic code。（ORNL）

Metaphase：分裂中期。在有丝分裂或减数分裂中染色体排列在细胞中间平面的阶段。（ORNL）

Microarray：微阵列。许多小型化的化学反应区域，可以用来检测 DNA 片段、抗体或蛋白质。（ORNL）

Microbial genetics：微生物遗传学。一门研究细菌、古生菌以及其他微生物的基因或基因功能的学科，常用于生物治疗、可用能源以及疾病防治等领域的研究。参见 model organisms、biotechnology、bioremediation。（ORNL）

Microinjection：显微注射。一种利用极细的微毛细管将外源 DNA 注入细胞的技术。（ORNL）

Mitochondrial DNA：线粒体 DNA。线粒体中的遗传物质（线粒体能为细胞产生能量），与细胞核 DNA 的遗传方式不同。参见 cell、DNA、genome、nucleus。（ORNL）

Mitosis：有丝分裂。细胞中细胞核分裂产生子细胞的过程，每个子细胞完全相同，并且与亲本细胞完全相同。参见 meiosis。（ORNL）

Modeling：建模。指使用统计学分析、计算机分析或利用模式生物来对研究成果进行预测。（ORNL）

Model organisms：模式生物。指用于研究的实验动物或其他生物。（ORNL）

Molecular biology：分子生物学。一门研究生物学中重要分子的结构、功能和组成的学科。（ORNL）

Molecular farming：分子作物农场。指发展转基因动物来生产医用的人类蛋白。（ORNL）

Molecular genetics：分子遗传学。一门研究生物遗传中重要的大分子的学科。（ORNL）

Molecular medicine：分子医学。在分子水平对损伤或疾病的治疗。分子医学的例子包括，基于 DNA 的诊断检测或根据 DNA 序列信息制药。（ORNL）

Monogenic disorder：单基因病。指由单基因突变而引起的病症。参见 mutation、polygenic disorder。(ORNL)

Monogenic inheritance：单基因遗传。参见 monogenic disorder。(ORNL)

Monosomy：单倍体。仅拥有单套染色体组而不是通常的两套染色体组。参见 cell、chromosome、gene expression、trisomy。(ORNL)

Morbid map：致病图。用来显示与疾病相关的基因在染色体上的位置的图表。(ORNL)

Motif：序列模式。(a) 蛋白质序列中短的保守区域。它们是结构域中保守性很高的部分。(NCBI)。(b) 多肽链上短的保守区域。序列模式相同并不代表是同系物。(SMART)

Mouse model：小鼠模式。参见 model organisms。(ORNL)

Multifactorial or multigenic disorder：多基因病。参见 polygenic disorder。(ORNL)

Multiple sequence alignment：多序列比对。三个或多个序列之间的比对，如果序列在同一列有相同结构位置的残基和（或）祖传的残基，则会在该位置插入空位。ClustalW 是一种最为广泛使用的多序列比对程序之一。(NCBI)

Multiplexing：复合技术。一种测序手段，可同时对几组样本测序以加快测序速度。(ORNL)

Murine：鼠科动物。在 *Mus* 类的生物体。鼠或小鼠。(ORNL)

Mutagen：诱变剂。一种能够引起细胞的基因持久变化的试剂。不包括发生在正常的基因重组时的变化。(ORNL)

Mutagenicity：诱变性。能引起持久性基因改变的化学或物理试剂所具有的性质。参见 somatic cell genetic mutation。(ORNL)

Mutation：突变。DNA 序列的可以遗传性变化。参见 polymorphism。(ORNL)

N

Nitrogenous base：含氮碱基。一种含氮的分子，具有碱基的化学特性。DNA 包含的含氮碱基有腺嘌呤 (A)、鸟嘌呤 (G)、胞嘧啶 (C) 以及胸腺嘧啶 (T)。参见 DNA。(ORNL)

Northern blot：Northern 印迹。将特异的 RNA 从琼脂胶转移至硝酸纤维素滤膜上与互补 DNA 序列杂交的分子技术。参见 DNA、library。(ORNL)

Nuclear transfer：细胞核移植。将一个细胞的细胞核移植至一个细胞核已经移除的卵细胞中，这样来自供体细胞的遗传信息会控制细胞核移植所形成的细胞。这样的细胞可以被诱导形成胚胎。这个过程曾用来制造克隆羊 Dolly。参见 cloning。(ORNL)

Nucleic acid：核酸。由核苷子单位组成的大分子。参见 DNA。(ORNL)

Nucleolar organizing region：核仁组织区。染色体中核糖体 RNA 基因 (rDNA) 所在的部位。(ORNL)

Nucleotide：核苷酸。DNA 和 RNA 的一种结构成分或组成。一个核苷酸包含一个含氮碱基（在 DNA 中是腺嘌呤、胸腺嘧啶、鸟嘌呤和胞嘧啶；在 RNA 中是腺嘌呤、胸腺嘧啶、鸟嘌呤和尿嘧啶)、一个磷酸分子，以及一分子的糖 (DNA 中是脱氧核糖，RNA 中是核糖)。数千个核苷酸分子连接组成了一个 DNA 或 RNA 分子。参见 DNA、base pair、RNA。(ORNL)

Nucleus：细胞核。真核细胞中包含大部分遗传物质的细胞器官。(ORNL)

O

Oligo：寡核苷酸。参见 oligonucleotide。(ORNL)

Oligogenic：寡基因性状。由两个或多个基因共同控制的表型性状。参见 polygenic disorder。(ORNL)

Oligonucleotide：寡核苷酸。一种通常由 25 个或更少的核苷酸组成的分子；常用作 DNA 合成时的引物。参见 nucleotide。(ORNL)

Oncogene：致癌基因。一种与癌症相关的基因。许多致癌基因直接或间接地与控制细胞增殖速度有关。(ORNL)

Open reading frame (ORF)：开放阅读框。位于 DNA 或 RNA 上起始密码子与终止密码子之间的序列。(ORNL)

Operon：操纵子。一组在操纵基因的控制下进行转录的基因。(ORNL)

Optimal alignment：最佳联配。两个序列之间有最高打分值的排列。(NCBI)

ORF：开放阅读框。参见 open reading frame。(SMART)

Orthologous：直系同源。指不同种类的同源序列，它们是在物种形成事件中从一个祖先序列独立进化形成的；可能有相似的功能，也可能没有。(NCBI)

Overlapping clones：重叠克隆。参见 genomic library。(ORNL)

P

P value：*P* 值。在比对时，获得某个打分值或更高的打分值的可能性。通过数据库中具有相同长度或组分的随机序列之间的比对，可以得到高打分值的片段的预期分布，将它与观察到的比对打分值 *S* 相连，就可以计算出 *P* 值。显著性最高的 *P* 值应该接近于零。*P* 值和 *E* 值用不同的方法来表示比对的显著性。(NCBI)

P1-derived artificial chromosome (PAC)：P1 衍生的人工染色体。一种用于在大肠杆菌中克隆 DNA 片段的载体，其插入片段长约 100～300kb，平均长度约 150kb。它是基于噬菌体（一种病毒）P1 的基因组的。参见 cloning vector。(ORNL)

PAM：可接受点突变。一个用于衡量蛋白质序列的进化突变程度的单位。1 个 PAM 的进化距离表示蛋白质序列中平均 1%的氨基酸残基发生突变的概率。PAM (x) 替换矩阵是一个查找表，其中每个氨基酸残基的替换打分值是基于进化趋异程度为 *x* 的紧密相关蛋白的替换频率而计算的。(NCBI)

Paralogous：共生同源。指在单个种类中由于基因复制事件而产生的同源序列。(ORNL)

Partly open barrel：部分打开的桶状结构。其边缘上的分支链并没有完全的形成氢键，因为这些分支链中的一个在由多个残基连接的两部分中。这些边缘分支链可看作是单个但被中断的分支链，可以用有效的分支链数 $n*$ 和剪力参数 $S*$ 来分类。在很少有部分打开的桶状结构中，β 折叠仅仅由边缘分支链之间的少数旁链氢键所连接。(SCOP)

Patent：专利。在遗传学上指某个人或组织对基因、基因变异，以及已测序的遗传物质的可识别部分等的特有权利。参见 gene。(ORNL)

Pedigree：谱系。家族谱系图表，显示了显著的特性或疾病是如何遗传的。参见 inherit。(ORNL)

Penetrance：外显率。一个基因或遗传性状能够被表达的可能性。如果某个人群有控制某遗传性状的基因，完全外显率指该性状在整个人群中都能全部得以表达；不完全外显率指该性状只能在部分人群中得以表达。外显率的大小也可能随着人群中的年龄范围而改变。(ORNL)

Peptide：肽。指两个或两个以上的氨基酸，通过一个氨基酸的氨基和另一个的羧基结合而形成的键叫“肽键”。参见 polypeptide。(ORNL)

Phage：噬菌体。一种自然宿主为细菌细胞的病毒。(ORNL)

Pharmacogenomics：药物基因组学。一门利用个体的遗传信息获取最好的药物疗效，同时具有最小副作用的研究领域。(ORNL)

Phenocopy：拟表型。一种不是由基因遗传引起的性状，但是在表型上与遗传性状完全相同。(ORNL)

Phenotype：表型。生物体的物理特征或是否表现患有疾病。表现型的特征不是必定遗传的。参见 genotype。(ORNL)

Physical map：物理图谱。物理图谱描绘 DNA 上可以识别的标记的位置和相互之间的距离（以碱基对的数目为衡量单位），这些可以识别的标记包括限制性内切酶的酶切位点、基因等。物理图谱不考虑两个标记共同遗传的概率等信息。对于人类基因组来说，最粗略的物理图谱是染色体的条带染色模式，最精细的图谱是测出染色体的完整核苷酸序列。(ORNL)

Plasmid：质粒。质粒是细菌的染色体外能够自我复制的环状 DNA 分子。它能够和细胞核中的染色体明显地区别开来，而且并不是细胞生存的必要物质。一些质粒适宜于整合到宿主细胞中去，并利用宿主细胞的 DNA 大量繁殖。许多人工构造的质粒常用作克隆载体。(ORNL)

Pleiotropy：基因多效性。指造成多种表型（如多病综合征）的基因。(ORNL)

Pluripotency：多能性。指细胞所具有的分化成多种成熟细胞的潜能，它依赖于环境。(ORNL)

Polygenic disorder：多基因病。由多个基因位点共同决定的遗传病（如心脏病、糖尿病、一些癌症等）。这类疾病的遗传由多个基因位点共同控制，因而比单基因病的遗传更为复杂。参见 single-gene disorder。(ORNL)

Polymerase chain reaction (PCR)：聚合酶链式反应。一种体外扩增 DNA 的方法。PCR 使用一种耐热的多聚酶，以及两个含有 20 个碱基的单链引物。经过高温变性将模板 DNA 分离成两条链，低温退火使得引物和一条模板单链结合，然后是中温延伸，反应液的游离核苷酸紧接着引物从 5′端到 3′端合成一条互补的新链。而新合成的 DNA 又可以继续进行上述循环，因此 DNA 的数目不断倍增。PCR 技术也可以用来探测 DNA 样本中指定序列的存在。(ORNL)

Polymerase, DNA or RNA：DNA 或 RNA 聚合

酶。一种能促进核苷酸与模板核苷酸结合的酶，能将核苷酸组装成 RNA 或将脱氧核苷酸组装成 DNA。(ORNL)

Polymorphism：多态性。多个个体之间 DNA 的差异称为多态性。在种群中，DNA 变异概率超过 1% 的变异，比较适宜作为绘制连接图谱的证据。参见 mutation。(ORNL)

Polypeptide：多肽。蛋白质或蛋白质的一部分，由一串通过肽键连接的氨基酸组成。(ORNL)

Population genetics：群体遗传学。研究群体中不同个体之间基因差异的学科。(ORNL)

Positional cloning：定位克隆。一种用来确定基因的技术，通常根据基因在染色体上的位置来确定那些与疾病相关的基因。(ORNL)

Primer：引物。预先制备的比较短的核苷酸链，新的脱氧核糖核苷酸可以在 DNA 聚合酶的作用下结合在引物的后面。(ORNL)

Privacy：隐私。指在遗传学上，人们对自己的遗传信息有保密的权利。(ORNL)

Probe：探针。是一条单链 DNA 或者一条 RNA 链，具有特定的序列，并且使用放射性元素或免疫特性物质进行标记。探针和克隆库中的某条互补片段结合成一条双链结构，我们可以通过杂交并借助于探针的检测获知与其互补的碱基序列的位置。(ORNL)

Profile：表达谱。(a) 一种罗列了蛋白质序列的每个位置上每个氨基酸出现频率的表格。这些频率是通过对包含指定结构域的序列进行多次比对而得到的。参见 PSSM。(NCBI)。(b) 一个显示某个同源家族中的特定位置的打分值和空位罚分的表格，可以用于搜索序列数据库。在 ClustalW 衍生的表达谱中，关系越远的序列，其配分权重越高。(SMART)

Prokaryote：原核生物。一种细胞或生物体，没有细胞膜、结构清晰的细胞核以及其他细胞器。细菌是原核生物。参见 chromosome，eukaryote。(ORNL)

Promoter：启动子。DNA 上的一个特定位点，RNA 聚合酶在此和 DNA 结合，并由此开始转录过程。(ORNL)

Pronucleus：生殖核。受精前的精子或卵子的核。参见 nucleus、transgenic。(ORNL)

Protein：蛋白质。一种由一条或者多条肽链构成的大分子。每条肽链上核苷酸的顺序是由基因外显子部分的碱基序列决定的。蛋白质对机体的细胞、组织和器官的结构、功能及调节都有重要作用。每种蛋白质都具有特定的功能。酶、抗体和激素等都是蛋白质。(ORNL)

Proteome：蛋白质组。一个细胞或器官在特定时间、特定环境下所表达的蛋白质。(ORNL)

Proteomics：蛋白质组学。对某个生物体中正常或生病组织的蛋白质表达进行系统的分析，包括对所有蛋白的分离、鉴定和特征化。(NCBI)

Pseudogene：伪基因。类似于基因却没有功能的 DNA 序列；可能是以前的功能性基因经过多次突变后的残留物。(ORNL)

PSI-BLAST：特定位点迭代 BLAST 程序。使用 BLAST 算法进行迭代搜索。初始搜索产生的表达谱会在接下来的搜索中使用。这个过程会根据需要而一直重复，在每个循环中产生的新序列会用来重新定义表达谱。(NCBI)

PSSM：特定位点记分矩阵。PSSM 给出了在目标序列中寻找特定的相配对的氨基酸的对数比分值。参见 profile。(NCBI)

Purine：嘌呤。一种在核苷酸中出现的含氮、双环的基础化合物，在 DNA 中的嘌呤是腺嘌呤 A，而在 RNA 中的是鸟嘌呤 G。参见 base pair。(ORNL)

Pyrimidine：嘧啶。一种在核苷酸中出现的含氮、单环的基础化合物。在 DNA 中的嘧啶是胞嘧啶 C 和胸腺嘧啶 T，而在 RNA 中的是胞嘧啶 C 和尿嘧啶 U。参见 base pair。(ORNL)

Q

Query：检测。输入序列（或其他搜索项）与数据库中的所有条目进行的比较。(NCBI)

R

Radiation hybrid：放射性杂交。其杂交细胞包含受辐射的人类染色体的小片段。染色体上的辐射位点图谱为人类基因组、老鼠基因组、小鼠基因组及其他基因组提供了重要的标记，这可以构建非常精确的序列标记位点图谱，从而对多因子病进行研究。参见 sequence-tagged site。(ORNL)

Rare-cutter enzyme：参见 restriction enzyme cutting site。(ORNL)

Raw score：初值。指通过计算替换和空位所得打分值之和而得到的联配值 S。替换打分值以查找表的形式表示。空位打分值是通过计算空位开放罚分 G 和空位拓展罚分 L 求和而得到的。对于长度为 n 的空位，空位罚分值是 $G+Ln$。空位罚分 G 与 L 的

选择完全是根据经验，通常 G 选择一个较高的数值（10～15），L 选择一个较低的数值（1～2）。参见 PAM、BLOSUM。(NCBI)

Recessive gene：隐性基因。只有在纯合时才能表达的基因；但对于男性来说，X 染色体上的隐性基因也可以表达。(ORNL)

Reciprocal translocation：相互易位。指一对染色体相互交换长度和位置相同的 DNA 片段。这会造成基因的混乱。(ORNL)

Recombinant clone：重组克隆。指包含重组 DNA 分子的克隆。参见 recombinant DNA technology。(ORNL)

Recombinant DNA molecules：重组 DNA 分子。利用重组 DNA 技术而结合的不同来源的 DNA 分子的组合。(ORNL)

Recombinant DNA technology：重组 DNA 技术。在无细胞系统（细胞或生物体外）中将 DNA 片段结合在一起的技术。在适宜的条件下，将重组 DNA 分子导入宿主细胞进行自动复制扩增，或是在与细胞染色体结合后进行复制扩增。(ORNL)

Recombination：重组。通过基因结合产生后代的过程，与父母的遗传基因有所不同。在高等动物中，重组是通过交换发生。参见 crossing over、mutation。(ORNL)

Regulatory region or sequence：调节区或调节序列。控制基因表达的 DNA 碱基序列。(ORNL)

Repetitive DNA：DNA 重复序列。在基因组中有多个拷贝的不同长度的序列；人类基因组的大部分都是重复序列。(ORNL)

Reporter gene：报告基因。参见 marker。(ORNL)

Resolution：分辨率。DNA 物理图谱上分子细节的等级，范围从低到高。(ORNL)

Restriction enzyme cutting site：限制性酶切位点。一段特殊的 DNA 核苷酸序列，特定的限制性酶会切割这个 DNA 片段。某些位点在 DNA 上出现的频率很高（如每几百个碱基对出现一次），而另外一些频率低些（如每一万个碱基出现一次）。(ORNL)

Restriction enzyme, endonuclease：限制酶，限制性内切酶。一种能识别特殊的短核苷酸序列，并在 DNA 的某些位点上切割的蛋白质。细菌包含了 400 种这样的酶，能识别和切割 100 种以上不同的 DNA 序列。参见 restriction enzyme cutting site。(ORNL)

Restriction fragment length polymorphism (RELP)：限制性片段长度多态性。特殊限制酶切割不同个体出现的 DNA 片段大小不同的差异。导致 RFLP 的多态性序列在物理图谱和遗传连锁图上被用作标记。RFLP 通常是因切割位点发生突变所致。参见 marker、polymorphism。(ORNL)

Retroviral infection：逆转录感染。逆转录病毒载体（例如一些病毒）能够使用自己的重组 DNA 将本身的遗传物质插入到宿主细胞的染色体上。然后它们就可以通过宿主细胞进行繁殖。(ORNL)

Reverse transcriptase：逆转录酶。指逆转录病毒以自身的 RNA 为模板，合成互补 DNA（cDNA）时所使用的酶。生成的 DNA 会插入到宿主细胞的染色体上。(ORNL)

Ribonucleotide：核糖核苷酸。参见 nucleotide。(ORNL)

Ribose：核糖。一种五碳糖，RNA 的组分之一。参见 ribonucleic acid、deoxyribose。(ORNL)

Ribosomal RNA (rRNA)：核糖体 RNA。在细胞核糖体中发现的一种 RNA。(ORNL)

Ribosomes：核糖体。细胞质中含有特定的核糖体 RNA 和相关蛋白质的细胞器，是蛋白质的合成场所。参见 RNA。(ORNL)

Risk communication：风险沟通。在遗传学上这是一种教育的过程，在该过程中，遗传咨询师或其他私人医生试着解释遗传测验结果，并且针对结果向病人以及他们的后代提出建议。

RNA (ribonucleic acid)：从细胞的细胞核和细胞质中分离出来的化学物质，在细胞的蛋白质合成和其他生化反应中起着重要作用，RNA 的结构和 DNA 的结构类似。RNA 可以分为信使 RNA、转运 RNA、核糖体 RNA 以及其他类型的 RNA，每种都行使不同的功能。(ORNL)

S

Sanger sequencing：Sanger 测序。一种广泛使用的测定 DNA 碱基顺序的方法。参见 sequencing、shotgun sequencing。(ORNL)

Satellite：随体。指位于染色体末端的球形染色体节段，通过次缢痕区与染色体主体部分相连。(ORNL)

Scaffold：支架。由序列重叠群拼接而成。(ORNL)

Seed alignment：种子联配。指只含有每对同系物中的一个的比对；这些同系物在 ClustalW 衍生的系统树中用距离小于 0.2 的分支连接。(SMART)

SEG：一种过滤氨基酸序列中低复杂区域的程序。

在比对中被过滤掉的氨基酸用“X”表示。在BLAST2.0的blastp子程序中，SEG过滤是默认执行的。(NCBI)

Segregation：分离。遗传学上，在细胞减数分裂过程中，同源染色体的两个片段相互分离，并随机分配到配子细胞中。(ORNL)

Sequence：序列。参见base sequence。(ORNL)

Sequence assembly：序列组合。将多个已测序的DNA片段进行排序的过程。(ORNL)

Sequence-tagged site (STS)：序列标记位点。一段短的DNA序列(200～500bp)，这种序列在人类基因组中只出现一次，其位置和碱基顺序都是已知的。在PCR反应中可以检测出STS，STS对染色体位置的绘制和从不同实验动物中得到测序数据是十分有用的。STS在人类基因组的物理图谱上是作为标记使用的。表达序列标记(ETS)是cDNA上的STS。(ORNL)

Sequencing：测序。指确定DNA或RNA分子的核苷(碱基序列)顺序，或蛋白质的氨基酸顺序。(ORNL)

Sequencing technology：测序技术。用来测定DNA中核苷酸序列的装置和操作程序。(ORNL)

Sex chromosome：性染色体。在人类细胞中是X染色体或Y染色体，性染色体决定了个体的性别。雌性细胞中含有两个X染色体，而雄性细胞中含有1个X染色体和1个Y染色体。性染色体由染色体组型中的第23对染色体组成。参见autosome。(ORNL)

Sex linked：性连锁。与X染色体或Y染色体相关的遗传性状或疾病，通常只在男性中发生。参见gene、mutation、sex chromosome。(ORNL)

Shotgun method：鸟枪法。一种测序方法，包括从基因组中获得随机的、已测序的克隆片段；并且对初始基因的位置一无所知。鸟枪法与直接测序法(这种方法中测序的DNA片段在染色体上的位置是已知的)形成对比。这两种方法各有优点，研究者可以将二者结合，以用来对人类基因组进行测序。参见library、genomic library。(ORNL)

Similarity：相似性。指核苷酸或蛋白质序列的相关程度。两个序列之间的相似性是基于相同和(或)保守序列所占的百分比的。在BLAST中，相似性指一个正定的打分值矩阵。(NCBI)

Single-gene disorder：单基因病。由单个基因的等位基因发生突变而导致的遗传病(如杜兴肌肉营养不良、眼癌、镰刀形细胞症)。参见polygenic disorders。(ORNL)

Single-nucleotide polymorphism (SNP)：单核苷酸多态性。指由于基因组序列中单个核苷酸(A、T、C或G)发生改变而引起的DNA序列的差异。参见mutation、polymorphism、single-gene disorder。(ORNL)

Somatic cell：体细胞。个体中除了生殖细胞及其母细胞之外的细胞，都是体细胞。参见gamete。(ORNL)

Somatic cell gene therapy：体细胞基因治疗。将新的遗传物质植入细胞，使之表达基因产物，以达到治疗目的。植入的遗传物质不能遗传给后代。参见gene therapy。(ORNL)

Somatic cell genetic mutation：体细胞遗传突变。指不能遗传给后代的遗传结构的改变，也称为“获得性突变”。参见germ line genetic mutation。(ORNL)

Southern blotting：DNA印迹技术。指将基因组DNA进行凝胶电泳，再将胶中的DNA分子转移到过滤膜上，使用放射性互补探针来探测特定碱基序列的技术。

Spectral karyotype (SKY)：光谱染色体组型。将一种生物体的所有染色体形象化的方法，每条染色体将标上不同的颜色。这种技术对确定染色体非正常化是十分有用的。参见chromosome。(ORNL)

Splice site：剪接位点。指RNA在将DNA序列转录为一段连续的基因从而翻译成蛋白质时，会在从剪接位点开始把非编码区移除。(ORNL)

Sporadic cancer：散发性癌。指随机发生的并且不是从亲本遗传的癌。它由一个细胞中的DNA改变引起，随后生长、分化，并且在机体中扩散。参见hereditary cancer。(ORNL)

Stem cell：干细胞。骨髓中未分化的原胞，具有繁殖和分化成特定血细胞的能力。(ORNL)

Structural genomics：结构基因组学。指使用实验技术和计算机模拟来确定大量蛋白质的三维结构的学科。(ORNL)

Substitution：替换。(a)在指定的位置不相同的氨基酸进行联配。如果联配的残基有相似的物理化学性质，那么替换是保守的。(NCBI)。(b)遗传学的一种突变类型，用另一个核苷酸替代DNA序列中的核苷酸，或用另一个氨基酸替代蛋白质中的氨基酸。参见mutation。(ORNL)

Substitution matrix：替换矩阵。替换矩阵中的值

与氨基酸对中第 i 个氨基酸突变为第 j 个氨基酸的概率成比例。构建这样的矩阵需要组装一个大的、含有不同的成对排列的氨基酸样本。如果样本足够大，其统计性显著，那么得到的替换矩阵可以反映经过某一阶段进化后的突变概率的真实值。(NCBI)

Suppressor gene：抑制基因。能够抑制另一个基因活性的基因。(ORNL)

Syndrome：症候群。一种综合征或畸形的集合，显示了某种疾病的特性。

Syngeneic：同基因的。指同一种类中遗传性质相同的成员。(ORNL)

Synteny：共线性。指不同种类中基因在染色体上的顺序相同的现象。(ORNL)

T

Tandem repeat sequences：串联重复序列。染色体上相同碱基序列出现多个拷贝的现象，这些相同序列常用作物理图谱中的标记子。参见 physical map。(ORNL)

Targeted mutagenesis：靶点突变。指直接在染色体的特定位点上人为对遗传结构进行改变。常用于研究靶点区域的功能。参见 mutation、polymorphism。(ORNL)

Technology transfer：技术性转移。指科学发现从研究室转移到商业部门的过程。(ORNL)

Telomerase：端粒末端转移酶。一种指导端粒复制的酶。(ORNL)

Telomere：端粒。染色体的末端部分，这一特殊结构区域与线性 DNA 分子的复制和稳定相关。

Teratogenic：畸形因素。指造成胚胎畸形发育的物质（如化学药品或放射物）。参见 mutagen。(ORNL)

Thymine (T)：胸腺嘧啶。一种含氮碱基，是碱基对 A-T（腺嘌呤-胸腺嘧啶）的一部分。参见 base pair、nucleotide。(ORNL)

Toxicogenomics：毒理基因组学。研究基因对环境压力或有毒物质的反应的学科。使用生物信息学将全基因组的 mRNA 表达谱与蛋白质表达模式结合起来，从而了解基因与环境之间的相互关系在疾病和功能障碍中所起的作用。(ORNL)

Transcription：转录。以一条 DNA 序列（一个基因）为模板合成 RNA 的过程；它是基因表达的第一步。参见 translation。(ORNL)

Transcription factor：转录因子。一种结合调节区域并帮助控制基因表达的蛋白质。(ORNL)

Transcriptome：转录组。指特定时间特定组织上所有被激活的基因、mRNA 或转录物的总和。(ORNL)

Transfection：转染。将外源 DNA 引入宿主细胞的过程。参见 cloning、vector、gene therapy。(ORNL)

Transfer RNA (tRNA)：转移 RNA。一种具有与 mRNA 三核苷编码序列匹配的三联核苷序列结构的 RNA。转移 RNA 在蛋白质合成中的任务是，与氨基酸结合，并将其运载入核糖体中，在那里根据 mRNA 运送的遗传密码组装成蛋白质。(ORNL)

Transformation：转化。通过在基因组上整合外源 DNA 来改变个体细胞遗物质的过程。(ORNL)

Transgenic：转基因的。一种由实验产生的有机体，它的 DNA 被人工地导入并与生物体种系的 DNA 结合。参见 cell、DNA、gene、nucleus、germ line。(ORNL)

Translation：翻译。mRNA 携带的遗传密码指导氨基酸合成蛋白质的过程。参见 transcription。(ORNL)

Translocation：易位。破坏和移动某条染色体上的大片段，然后将其连接到不同的染色体上的过程。参见 mutation。(ORNL)

Transposable element：转座元件。指一类可以从一个染色体位点转移到另一个的 DNA 序列。(ORNL)

Trisomy：三（染色）体性。带有某个特殊染色体的三个拷贝，而不是通常的两个拷贝的现象。参见 cell、gene、gene expression、chromosome。(ORNL)

U

Unitary matrix：酉矩阵，幺正矩阵。也称为单位矩阵。是一个只有在字符相同时才能得到正打分值的打分系统。(NCBI)

Up and down：上下型。螺旋束或折叠叶的最简单的拓扑结构，其中连续的螺旋是相邻和反平行的；它近似等同于 β 折叠的回纹拓扑结构。(SCOP)

Uracil：尿嘧啶。一种含氮碱基，通常只存在于 RNA 而非 DNA 中；它能与腺嘌呤形成一个碱基对。参见 base pair、nucleotide。(ORNL)

V

Vector：载体。参见 cloning vector。(ORNL)

Virus：病毒。一种非细胞的生物学实体，只能在宿

主细胞中复制，病毒由核酸和蛋白质外壳组成；但一些动物病毒有膜包裹。在被感染的细胞中，病毒利用宿主的合成能力复制它的子代。参见 cloning vector。(ORNL)

W

Western blot：一种识别和定位蛋白质的技术，以蛋白质与特异抗体的结合能力为基础。参见 DNA、Northern blot、protein、RNA、Southern blotting。(ORNL)

Wild type：野生型。指某生物体在自然界中出现频率最多的形式。(ORNL)

Working draft DNA sequence：未完成的基因定序草图。参见 Draft DNA sequence。(ORNL)

X

X chromosome：X 染色体。两种性染色体（X 染色体和 Y 染色体）中的一种。参见 Y chromosome、sex chromosome。(ORNL)

Xenograft：异种移植物。指将一个种类生物体的组织或器官移植或嫁接到另一个种类（属，或家族）的生物体内。常见的例子是在人身体内使用猪的心脏瓣膜。(ORNL)

Y

Y chromosome：Y 染色体。两种性染色体（X 染色体和 Y 染色体）中的一种。参见 X chromosome、sex chromosome。(ORNL)

Yeast artificial chromosome (YAC)：酵母菌人工合成染色体。由酵母 DNA 所构建，它是一种能克隆长 DNA 片段的载体。参见 cloning vector、cosmid。(ORNL)

Z

Zinc-finger protein：锌指蛋白。一些包含锌原子的蛋白的次要特征；是一种结合 DNA 的蛋白。(ORNL)

表 1　因特网上的术语表

资　源	网　址
剑桥卫生研究院(Cambridge Healthtech Institute)的 RNA 术语表	http://www.genomicglossaries.com/content/RNA.asp
生物技术学科有关的术语表	http://smart.embl-heidelberg.de/help/smart_glossary.shtml
结构基因组学的术语表	http://scop.mrc-lmb.cam.ac.uk/scop/gloss.html
基因组术语表(美国橡树岭国家实验室，Oak Ridge National Laboratory)	http://www.ornl.gov/TechResources/Human_Genome/glossary/
美国国立人类基因组研究所(National Human Genome Research Institute，NHGRI)，带发声的遗传学相关术语表	http://www.genome.gov/glossary.cfm
NCBI 的 BLAST 术语表	http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html
剑桥卫生研究院(Cambridge Healthtech Institute)的基因组学术语表	http://www.genomicglossaries.com
Biotech 的生物学字典	http://biotech.icmb.utexas.edu/search/dict-search.html
CancerPage.com 的术语表	http://www.cancerpage.com/glossary
自然和加速生物恢复研究计划(Natural and Accelerated Bioremediation Research，NABIR)的生物治疗有关的术语表	http://www.lbl.gov/NABIR/generalinfo/glossary.html
多种语言的医学词汇	http://allserv.rug.ac.be/%7Ervdstich/eugloss/language.html
细胞和分子生物学词典	http://www.mblab.gla.ac.uk/dictionary

自测题答案

第 2 章

1 (b); 2 (c); 3 (a); 4 (d); 5 (a); 6 (a); 7 (c); 8 (d); 9 (c); 10 (c)

第 3 章

1 天冬酰胺 N; 谷氨酰胺 Q; 色氨酸 W; 酪氨酸 Y; 苯丙氨酸 F

2 (a); 3 (d); 4 (c); 5 (d); 6 (a); 7 (c); 8 错误; 9 (c); 10 (d)

第 4 章

1 (d); 2 (c); 3 (a); 4 blastp (d), blastn (a), blastx (c), tblastn (b), tblastx (e);

5 (c); 6 (a); 7 (a); 8 (b); 9 (b); 10 (c)

第 5 章

1 (a); 2 (b); 3 (b); 4 (c); 5 (b); 6 (a); 7 (a); 8 (a); 9 (b); 10 (d)

第 6 章

1 (c); 2 (c); 3 (d); 4 (a); 5 (c); 6 (c); 7 (a); 8 (b); 9 (b); 10 (c)

第 7 章

1 (c); 2 (c); 3 (a); 4 (b); 5 (d); 6 (d); 7 (a); 8 (d); 9 (a)

第 8 章

1 (a); 2 (c); 3 (b); 4 (c); 5 (c); 6 (b); 7 (d); 8 (b); 9 (a); 10 (c)

第 9 章

1 (a); 2 (c); 3 (d); 4 (c); 5 (d); 6 (c); 7 (a); 8 (b); 9 (d)

第 10 章

1 (b); 2 (d); 3 (c); 4 (b); 5 (b); 6 (b); 7 (c); 8 (a); 9 (a)

第 11 章

1 (d); 2 (b); 3 (c); 4 (a); 5 (b); 6 (a); 7 (b); 8 (a); 9 (a); 10 (c)

第 12 章

1 (c); 2 (a); 3 (d); 4 (c); 5 (a); 6 (a); 7 (d)

第 13 章

1 (c); 2 (d); 3 (d); 4 (a); 5 (d); 6 (b); 7 (d)

第 14 章

1 (b); 2 (d); 3 (a); 4 (c); 5 (c); 6 (d); 7 (a)

第 15 章

1 (a)；2 (c)；3 (c)；4 (c)；5 (b)；6 (d)；7 (c)；8 (b)；9 (a)；10 (a)

第 16 章

1 (c)；2 (c)；3 (b)；4 (d)；5 (a)；6 (a)；7 (a)；8 (a)；9 (c)

第 17 章

1 (c)；2 (c)；3 (a)；4 (a)；5 (a)；6 (b)；7 (b)；8 (d)

第 18 章

1 (a)；2 (a)；3 (b)；4 (c)；5 (a)；6 (b)；7 (c)；8 (a)

索 引

主题索引

Alu 85
BLAST 67，80，441
blastn 127
BLOSUM62 93
BLOSUM 打分矩阵 55
CASP 297
CATH 266
cDNA 文库 152，156
ClustalW 327
ClustalX 327
CpG 岛 601
C 值矛盾 517
Dali 266
Dayhoff 模型 45
DDBJ 85
DNA 微阵列 163
DUST 87
EMBL 85，440
Ensembl 582
Ensembl 比对 120
Entrez 26，272
EST 85，153，157
eukaryotes 514
ExPASy 28，266
FASTA 124
FASTA 格式 82
GAP 程序 92
GenBank 13，82，84，85
Genetics Computer Group 63
GLIMMER 460
GSS 85
hidden Markov models（HMMs） 104
HIV-1 *pol* 99
HIV-1 pol 81
HIV 抗药性数据库 435
HSPs 93
HTGS 85
igBLAST 122
k 近邻聚类算法 194
LANL 435
MUMmer 471
NCBI 20，80，266，434，582
Needleman-Wunsch 80
Needleman-Wunsch 算法 58
nr 数据库 84
PAM1 矩阵 49
PDB 85
PDB 数据库 266
Pfam 105，315，439
PIR 84
PRF 84
PSI-BLAST 86，441
RBP4 99
RefSeq 82
REPBASE 85
RepeatMasker 530
Rosetta Stone 245
Rosetta Stone 模型 296
SCOP 266
Sim4 127

SMART 105
Smith-Waterman 80
Smith-Waterman 算法 64
STS 85
SwissProt 84
TaxPlot 470
tblastn 122
TIGR 120
UCSC 592
UniGene 26
UPGMA 方法 364
WU 比对 2.0 算法 124
白腐真菌 506
白念珠菌 504
保守结构域数据库 86，321
比对搜索 118
表达序列标签 17
不确定性理论 346
超家族 281
串联重复序列 523
从头预测 293
单核苷酸多态性 635
蛋白质比对 37
蛋白质结构域 281
蛋白质聚类工具 426
蛋白质数据银行 84
动态规划 62
非编码 RNA 531，606
非整倍体性 640
非转录区 81
分级鸟枪法 597
分子时钟假说 343
分子条形码 502
冈比亚疟蚊基因组 562
公共数据库 9
功能基因组学 496
功能基因组学 261
孤独症 639
海鞘 563
河豚 563
黑腹果蝇基因组 560
后生生物 558
基因表达系列分析 158
基因复制机制 491
基因预测软件 534
基因组测序 405
基因组测序计划 388
基因组分析 401
基因组学 3
基因组注解 410
基质辅助激光解析-电离飞行时间质谱 238
家族 281
间隙 43
检测限度 57
简并序列 424
简单分子构架研究工具 321
渐进比对 305
结构基因组 261
结构基因组学 262
结构域 118
进化树 553
局部比对 58
开放阅读框 460
雷特症 626
类 281
类病毒 420
连接近邻法 368
裂殖酵母 449，506
灵长类动物基因组 568
轮廓 223
脉孢菌 505
孟德尔疾病 621
模式 215
模式识别 BLAST 134
模体 213
拟南芥基因组 553
逆转录病毒 432
逆转录致癌基因 104
黏菌 557
疟原虫 547
欧洲生物信息学研究所（EBI） 28
皮尔逊相关系数 192
啤酒酵母 482
气味分子结合蛋白（OBP） 100
签名 215
趋同进化 424
全基因组测序 515
全局比对 58

全局归一化 180
染色体 538
染色体组芯片 641
人类基因组计划 581
融合 86
朊病毒 420
散点分析 181
伤残调整生命年 623
β-肾上腺素能受体 104
生物学数据库 12
视黄醇结合蛋白 6
视紫红质蛋白 104
噬菌体 421
疏水性绘图 226
双序列比对统计显著性 69
水稻 555
水平基因转移 461
丝氨酸蛋白酶抑制剂结构域 86
算法和程序 41
索引号码 22
同源模建 293
同源性 38
推测统计学 177
外源转座子 500
α-1-微球蛋白 86
微阵列 166
微阵列数据库 170
位点特异数据库 634
位点特异性反复比对 127
物种树 354
系统发生分析 355
系统发生树 348，384，442，464
系统发育分析 432
系统分类学 384
细菌和古细菌 446
显著性分析 190
线粒体遗传特性 643
小孢子寄生虫 504
小鼠基因组 564
小鼠基因组测序 565
新基因 490
秀丽隐杆线虫 559
序列标签位点 19
寻找基因 532
烟曲霉 504
遗传图和物理图 602
隐马模型 308
原核基因组 459
原生生物 544
孕激素相关子宫内膜蛋白 101
载脂蛋白家族 263
折叠子 215，281
真核基因 DNA 预测算法 533
真核基因组数据库 556
真核生物 514
真核生物 DNA 比较 541
真菌 480
脂质运载蛋白 86
直系同源 304
直系同源基因 619
植物基因组 549
重复 215
重复序列 602
主成分分析 200
转录因子数据库 538
转录组 147
转座子 499
转座子元件 520，545
自组织图算法 199
最大吝啬法 370

作者索引

Adrian Hill 650
Altschul 143
Altschul 和 Gish 94
Alvis Brazma 168
Anderson 392
Aparicio 563
Archibald Garrod 620
arren Gish 115
Ashburner 236
Baldauf 514
Barbara McClintock 520
Bateman 315
Benson 13
Blattner 516
Boeckmann 17
Bonaldo 156
Braunitzer 617
Brown 457
Carlton 549
Carl Woese 385
Chambaud 98
Churchill 165
Claverie 538
Cyrus chothia 133
Czelusniak 370
Da-Fei Feng 305
David Botstein 499
Davison 426
Dayhoff 52
Delcher 460
Dietmann 283
Dixon 104
Douglas 551
Downward 104
Dunham 599
Durbin 559
Dwight 486
Eddy 527
Eisen 461
Emile Zuckerkandl 73
Enright 245
Felsenstein 373
Ferretti 98
Fleischmann 394
Flores 420
Fox 454
Friedberg 134
Gardiner-Garden 和 Littlejohn 170
Gardner 547
Gaschen 435
Gavin 243
Gerstein 521
Gerstein 和 Levitt 262
Giaever 502
Glockner 558
Goff 556
Goffeau 397
Gotoh 66
Greenbaum 164
Hahn 432
Hall 553
Hastie 和 Bishop 153
Hendrix 424
Henikoff 323
Henikoff 和 Henikoff 55
Hillis 373
Hingamp 28
Ho 243
Holm 261
Huelsenbeck 373

IHGSC 595
Jacq 233
Jerzy Jurka 524
Johnson 和 Gibbs 420
Jukes 和 Cantor 341
Karlsson 109
Kellam 441
Kent 592
Kimura 341
Knight 517
Kohne 519
Koonin 461
Krogh 308
Kumar 532
Linus Pauling 73
Lipshutz 166
Luscombe 10
Makarov 533
Marco Pagni 和 C. Victor Jongeneel 115
Margaret Dayhoff 45
Margulies 149
Marti-Renom 293
Matsuda 617
Mayor 541
McClure 429
McGeoch 425
McGuffin 和 Jones 267
Mewes 482
Milner 和 Sutcliffe 148
Molloy 和 Witzmann 236
Morgenstern 331
Morrison 545
Motoo Kimura 347
Murzin 281
Myers 和 Miller 66
Needleman 和 Wunsch 58
Needleman 和 Wunsch 48
Nei 361
Newcomer 45
Newcomer 和 Ong 6
Ohyama 393
Oliver 394
Pagni 和 Jongeneel 94
Park 133
Patrick Brown 163
Pearson 和 Lipman 124
Poole 430
Pruitt 23
Prusiner 420
Ratjen 和 Doring 297
Reese 561
Richard Owen 446
Rose 531
Ross Macdonald 500
Rost 264
Russell Doolittle 305
Saitou 和 Nei 368
Sanger 391
Sankoff 490
Sasaki 537
Schaffer 134
Schmitt 155
Schuchhardt 180
Schultz 321
Sedgewick 41
Sigrist 220
Simons 296
Simpson 340
Smith-Waterman 64
Stauffer 568
Stein 410
Stephen Altschul, David Lipman 115
Suarez 423
Tamas 453
Tatusov 465
Thompson 331
Thornton 261
Upton 440
Velculescu 158
Venter 580
Walter M. Fitch 39
Waterston 564
Westbrook 17
Wheeler 23

William Pearson 115
Wilson 385
Winzeler 502
Woese 和 Fox 454
Wolfe 和 Shields 491
Wootton 和 Federhen 134
Wright 539
Zhang 144
Zuckerkandl 和 Pauling 343

［Ｇｅｎｅｒａｌ Ｉｎｆｏｒｍａｔｉｏｎ］
书名＝生物信息学与功能基因组学
作者＝［美］乔纳森　佩夫斯纳著　孙之荣主译
页数＝７０６
ＳＳ号＝１１６９４９６５
出版日期＝２００６
出版社＝化学工业出版社